T E S I D I D O T T O R A T O

Dipartimento di Scienze Statistiche e Matematiche "Silvio Vianelli"

Student Evaluation of Teaching: multilevel IRT model

La valutazione della didattica nell'opinione degli studenti: modelli multilivello IRT

Clara ROMANO

Tutor: *Prof.ssa Vincenza Capursi*

Coordinatore Dottorato: *Prof. Marcello Chiodi*

Dottorato di Ricerca in "Statistica, Statistica Applicata e Finanza Quantitativa", XXIII Ciclo - 2011
Settore Scientifico Disciplinare: SECS/S05 - Statistica

Università degli Studi di Palermo

To Tony and Serena

Acknowledgements

I want to thank all the people who have supported and motivated me in writing this work. I would like to thank prof. Vincenza Capursi for supervising this thesis, for advice and encouragement during my PhD course.
Special thanks are due to prof. Vito Muggeo for his suggestions. Other people have, in different ways, helped me through. In particular, special thanks go to dr. Salvatore Marcantonio, laudable fellow traveler with which I could report describes the topics of the thesis and dr. Miriam Tagliavia.
Also, I would also like to thank all the members of the Ph.D. program Board for their helpful comments and suggestions. I would like to extend my gratitude also to my colleagues and friends and all the members of the Department of "Statistical and Mathematical Sciences S. Vianelli" who contributed to make the past three years greatly enjoyable.
Finally, I would like to express my special appreciation to my family, my husband and my daughter for their patience while I had to organize my time between them and my studies. My gratitude goes to my parents and my parents in law for their care qand sacrifice which have always allowed me to study and complete my thesis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The assessment of teaching plays an important role within the Italian university. In 1993, the norm 537/93 established a national and local university evaluation system having the duty to monitor the productivity of research and of teaching. In accordance with this norm, Italian universities created theirNV (Nucleo di Valutazione) and began to operate in the wide and complex set of assessment activities. This has contributed to the collection of material relating to different aspects: the evaluation of university research, Students Evaluation of Teaching (SET), the analysis of student career by the universities.

The presence of a rating system has encouraged the formalization of "quality procedures" that are based on an internal and an external activity. The former refers to the process of maintaining and improving the quality, the latter to periodic quality assessment. A meaningful assessment, internal or external, requires the collection of points of view of those who participate in an important way to the training process. So it is necessary to take into account not only the opinions of professional trainers (teachers at various

levels and roles) but also students' opinions. The SET takes into account students' opinions through an evaluation form (questionnaire): this contains a set of compulsory items for all universities and it is structured in sections concerning several aspects of university courses (teaching, management aspects, class facilities, etc...)

The main purpose of this study is to measure the concept of quality of teaching in the opinion of students at the University of Palermo. This concept is considered a latent variable that cannot be directly observable and measurable. In order to measure the latent variable the construct needs to be operationalized in terms of a certain number of dimensions, which are measured through a set of indirect variables or a battery of items. In the psychometrical and psychological literature, multi-item Likert type scales are the main tools for measuring an underlying theoretical concept, which is not directly observable. So we can assess the quality of university courses obtaining an approximation of the true measure by indirect measurements provided by students' ratings.

Rating is an indicator of the level of the specific attribute that it is supposed to be measured by that item. Obviously, it is necessary to emphasize that the result depends on subjective factors, since each student is influenced by his/her own needs and expectations, which on the other hand depend on the different cultural backgrounds and the different socio-economic conditions. Apart from the methodological nature of the results, we should not forget that the point of view from which the evaluation originates is the opinion of students. So, due to the presence of heterogeneity of the opinions of students, you cannot expect to arrive at a measure of the quality of teaching based on a system of shared values.

1.1 History of SET

During the past few decades Student Evaluation of Teching (SET) has been considered as an important tool in the improvement of teaching quality even if Marsh (Marsh, 1984) and Wachtel (Wachtel, 1998) report that student evaluation programs were introduced to Harvard in 1915, and the first studies on SET effectiveness were written in the 1920s by Remmers (Remmers and Brandenburg, 1927; Remmers, 1928, 1929). Student evaluation research had a wide development in the 1970-1980 decade, when most of the research was devoted to the utility and validity of students' evaluation (Centra, 1993). Kulik (Kulik, 2011) states that the initial aim of SET served two goals: mapping the quality of teaching in universities, and providing information and help instructors in order to improve their teaching. For Marsh (Marsh, 1984) students ratings are also very useful to make administrative decisions and to satisfy a fundamental principle of the evaluation: the accountability. Although the implementation of SET was spread in many faculties, a lot of universities were resilient to the use and the utility of these ratings. Supporters argue that evaluative judgements have a strong positive influence on the improvement of instructional skills. Marsh (Marsh, 1987) states that opinions about the role of SET vary from "reliable, valid and useful" to "unreliable, invalid and useless". Today, more than 90% of U.S. universities use some sort of student evaluation mechanism to assess teaching (Murray, 2005). The desire to implement a measurement of teaching effectiveness based on student feedback is understandable and commentable. Students are one of the consumer groups interested in the product of an university education; therefore, their opinions are a vital source of information concerning the quality of instruction at institutions of higher education (Wright, 2006).

1.2 Aims

The general aim of this thesis is to measure the quality of teaching, through levels of satisfaction of students, on several aspects of university courses (items) i.e, students' ratings. The first question we asked is: what are the items that explain the overall students satisfaction? The second question is this: there are variables, such as student characteristics, that may influence the overall satisfaction? In the first phase of our work, we considered aggregated data in order to give some suggestions to the policy makers on the variables or items that determine the students' opinion. The statistical unit is the single teaching course. The purpose is to simplify the questionnaire, to give a policy making tool for the planning and for the improvement of the teaching. In particular we are interested to determine an explicit quantification of the relative importance of each item for the overall satisfaction of teaching that is a proxy of the teaching quality.
In the second phase of this work we want to analyze individual data (student), in order to take into account the students characteristics as variables within the model and to assess whether students' characteristics can affect the teaching evaluation. To sum up, the focal point of the thesis is the transition from an aggregated view of data to an individual view. Initially, using simple statistical tools, we tried to highlight any differences in terms of satisfaction among students. Subsequently, we applied more complex models in order to take into account the complex structure of our data and the student as the statistical unit. This was finalized to the introducion of the student characteristics as variables within a single model and to obtain more specific results.

1.3 Outline of the thesis

Chapter 2 is devoted to a descriptive analysis of the data. In particular we present descriptive and explorative analysis of the data. In particular, we also describe the main features of survey and the instrument of measurement adopted to reveal students' opinions.
In Chapter 3 we introduce the methodology for aggregated analysis. First of all we introduce the Indicator of Student Performance that combining age and UEC (University Educational Credits) give us information on students career. The indicator allows to split students in *bad* and *good* according to their performance. Subsequently, in order to determine what the items that explain the overall satisfaction for *bad* and *good* students are, we consider a regression linear model, in which items are covariates and the overall satisfaction item is the response variable. Various strategies can be adopted to deal with the previous issue. Our interest is to investigate the suitability of relative metrics in linear regression (Feldman, 2006; Grömping, 2007) as analitical tools for observational studies with correlated regressors.
Chapter 4 introduces the methodological framework underlying the Rasch model and its generalizations are introduced. The wide family of Multilevel (or Generalized Linear Mixed or Random) Models represents a methodological framework within which tha main part of IRT (Item Response Theory) models may be placed. The most famous application of the IRT approach has been proposed by the mathematician George Rasch in 1960 (Rasch, 1960) and it is known as Rasch model.
In the last decades a number of item response models have been developed as extensions of the Rasch model in the statistics and psychimetrics literature for the analysis of dichotomous and polytomous discrete responses: the Nominal Response Model (Bock, 1972), the Graded Response Model

(Samejima, 1969), the Rating Scale Model (Andrich, 1978), the Partial Credit Model (Masters, 1982). The most interesting part of these extensions concerns the structural part of the model and the effect of the predictors (students characteristics), which can be either fixed or mixed.
In Chapter 5 we present Multilevel IRT model results. Summarizing the focal aspects of our work, in Chapter 3 we consider aggregated data in linear regression model in which item C2 (that expresses the overall satisfaction declared by students) is a proxy of the quality of teaching; in Chapter 4 (and then in Chapter 5) we consider individual data in multilevel IRT model. Here items are at grade; in this way we obtain for which items students are more satisfied and so which items are the drivers of the quality of teaching.

Chapter 2

The survey

This study is based on data which are collected at a Faculty of the University of Palermo, from classes attending the academic year 2006-2007. Data on courses evaluations are provided by the Center for the Evaluation of University Activities, which is responsible for coordinating the survey on students' opinion about the quality of teaching at university.
The measurement instrument is an ad hoc questionnaire addressed to reveal students' opinion of course quality. The purpose is to assess the quality of university courses obtaining an approximation of the true measure by indirect measurements provided by students' ratings.

2.1 The survey plan

The plan for the detection of the opinion survey of students on the campus of Palermo can be summarized as follows:

a) target population: students who attend classes;

b) scope of the survey: the single teaching;

c) measurement instrument: questionnaire outlined by the Academic Senate;

d) the time of detection: the detection takes place during the last weeks of the terms.

In particular:

a) the reference population consists of students in classroom who take part in the questionnaire. So this is not a sample survey in a probabilistic sense, but a partial survey, as it is intended for students in attendance at that particular lesson;

b) as regards the object of detection, i.e. teaching to evaluate, it should be noted that all ongoing teaching should be detected. Indeed the coverage of the teachings evaluated, is not total. This is due mainly to the fact that financial resources by the university are not adequate;

c) a description of the measurement instrument is shown in the following paragraph;

d) the detection is performed only after the students have carried out at least three quarters of the total hours provided for each course. Moreover, classroom with less than 10 students were not considered.

What I have just said brings out the character of our cross-sectional observational study. In fact, neither study subjects nor the variables of interest (ie the items of the questionnaire) are manipulated by the researcher, you do not know in advance the characteristics of the subjects, the policy underlying the realism (Kish, 1987). Moreover, since the detection is made

at a precise moment in time we talk about cross-sectional study in which subjects have in common the fact that they attend the same course.
The questionnaire was administrated by 200 detectors, recruited among students of all faculties. After the survey, the questionnaires are sent to Centre for Evaluation where answers are trasferred on a computer via an optical reader. Finally, the data (aggregated by faculty and university) are sent to the NV of the University, which provides validation of their ”formal” analysis of the results and elaborates the final report.

2.2 The measurement instrument

The evaluation form used in the survey is structured in six sections. These sections provide information about students' personal details and students' opinion on several aspects of university courses, such as courses facilities, curriculum programming and teaching activities of the whole course. The preliminary section contains general information about the course (course code, type of degree, term, etc). The first section refers to student's personal characteristics (date of birth, residence, age, secondary school, number of credits collected, etc). Sections (B, C, D, E, F) contain items concerning various aspects of the course: teaching characteristics (B), global satisfaction and previous knowledge of the topic (C), management aspects (D), class facilities (E), teacher's characteristics (F). Finally, section G refers to courses organized in modules.
The items are measured on four categories according to the Likert scale: *definitely no*, *more no than yes*, *more yes than no*, *definitely yes*. Items B2, B6, B7, E2, have not considered in this study. In fact, in the faculty chosen for the survey the evaluation of tutorials, laboratory activities are not given

from all courses. For the same reason section G has not been considered. The item C1 was not considered because it is inherently 'bearer of quality', in the sense that it is our opinion that the interest in a discipline may positevely affect the assessment probably upwards, regardless of the intrinsic quality of provided service.
B9 is an ambiguous item. This refers to the level of previous knowledge on the topic with the intent to understand the contents of the course properly. The exclusion of the item B1 is motivated by the decision to eliminate all the questionnaires which have a percentage of less than 50%.
The removal of the item F1, for the percentage of classes taught by teachers owner, finds its reason in the difficulty of interpretation of the item itself. It is assumed that the high percentage of classes conducted by the teacher can be considered an enrichment of the concept of education?(in this case the item would be oriented positively with the quality of teaching)
The item B5 ("The teaching content is overlaid on his other teachings?") is also deleted: from previous analysis (Sulis, 2007), it was found that students interpret (you do not understand why) in a positive way the overlapping with other teachings.

2.3 Data

In our analysis we consider the undergraduate courses because they are more established and attended by more of students. The dataset consists of 8503 questionnaires, corresponding to 286 courses in the only undergraduate courses. The number of students per course range from 10 to 108 with a mean of 42 and a median value of 40. As table 2.1 shows, around the 15.3% of the questionnaire are the evaluation of courses with less than

20 evaluators. Whereas 68% of the courses have collected more than 30 evaluators; 156 courses out of 286 are medium classes, with a number of evaluators between 10 and 30; 108 classes have a number of evaluators between 31 and 60; 22 are large classes with more than 60 students; more than 100 questionnaires have been gathered only for one course. Summarizing, the largest classes (37.8%) are those with a number of students between 61 and 100, but the percentage of classes of small-medium size (33.2%) is also high.

n. stud. per class	n. students (%)		n. courses (%)	
10-20	1305	(15.34)	95	(33.22)
21-30	1444	(16.98)	61	(21.33)
31-60	4331	(50.93)	108	(37.76)
61-100	1326	(15.59)	21	(7.34)
>100	97	(1.14)	1	(0.35)
Total	8503	(100.0)	286	(100.0)

Table 2.1: Number of questionnaire per class

2.4 Students' characteristics

In this section we show some descriptive statistics of the characteristics of the respondents. The distribution of the gender variable (Table 2.2) shows a significant male presence in the faculty considered (76.8% of students). As far as the secondary school of origin is concerned, it can be noted that 63.9% of the questionnaires filled out by students come from high school. The percentage is much lower for students from other schools.

The univariate distribution of residence variable shows a majority of questionnaires completed by students in site (41.5%) compared to permanent students (37.2%).
A lower value (21.3%) is detected for commuter offsite students. Most of the students involved in the survey does not carry out any job (82.2%) and is not relevant to include students who have a full time job (2%)(Table 2.2).

modalities	n. students	% students
Gender		
Male	6289	76.77
Female	1903	23.22
Total	8192	100.0
Secondary School		
High school	5284	63.84
Other	2992	36.15
Total	8276	100
Residence		
in site student	3191	41.46
permanent resident student	2864	37.21
commuter offsite student	1642	21.33
Total	7697	100.0
Occupational Status		
no job	6771	82.22
part-time	1301	15.79
full-time	167	2.02
Total	8239	100.0

Table 2.2: Univariate distributions of Students' characteristics

The age distribution (Table 2.3) shows that questionnaires were filled out by students aged 19 to 21; just under 8% are over 24. The analysis of the characteristics concerning students' university *curricula* reveals that 80.8% are in course student ("regular student"), but more than half of the total of the questionnaires refers to students who have gathered less than 60 credits. The distribution of the number of credits (Table 2.3) already gathered by the student when he/she fills in the questionnaire is strongly skewed towards the bottom, with just 5.9% of the evaluation forms fulfilled by students who have gathered more than half of the credits.

2.5 Students' ratings

In this section we will analyze the distributions of ratings given by students. The students' ratings are measured by means of an ordinal scale with four categories. Since there is no information about distances between categories, as generally happens when working with ordinal scales, we prefer to avoid the attribution of scores and perform the analysis with appropriate statistical tools available for the type of variables. Table 2.4 shows frequency distributions (percentage) for each category of the 16 ordinal items. Almost all item distributions are positively skewed. In fact more than 50% of students gives positive responses to each items. In particular, items F2, F3, F4, F5 have median in the last category. Other items register the highest percentage of units in the *more yes than no* category. If items are ordered according to the percentage of students who are very satisfied, items F5, F3, F2, F4, F7 are in the first five ranking positions. At the bottom we find items E1, B10, B11, D1, D2 that concern managements aspects and coordination among courses. The last column of the table 2.4 shows an indicator that

modalities	n. students	% students
	Age	
18	369	4.34
19	2212	26.06
20	2099	24.73
21	1638	19.30
22	959	11.29
23	561	6.61
24	266	3.13
25	159	1.87
>26	225	2.65
Total	8488	100.0
	Number of credits collected	
0-30	3787	44.51
30-60	1759	20.69
60-90	1284	15.10
90-120	1003	11.79
120-150	505	5.93
150-180	165	1.94
Total	8503	100.0
	Regularity	
out of course student	1430	17.07
in course student	6765	80.75
repeating student	182	2.17
Total	8377	100.0

Table 2.3: Other students' characteristics

summarizes the students' ratings taking into account their heterogeneity (Bernardi *et al.*, 2004; Capursi and Librizzi, 2007). The general expression of the indicator is the following:

$$IS_{0.5} = 1 - \left(\frac{1}{k-1} \sum_{m=1}^{m-1} F_m^r \right)^{1/r}, \tag{2.1}$$

where F_m is the cumulative distribution function of items responses in correspondence to the modality m of the ordinal variable.
The 2.1 is a average power of order r. The average takes into account of the judgements variability. With the same average level of the distribution and symmetric distributions, when the variability of the distribution increases, the average increases if $r > 1$ and decreases if $r < 1$. The final expression is

$$IS_{0.5} = 1 - \left(\frac{1}{3} \sum_{m=1}^{3} F_m^{0.5} \right)^2, \tag{2.2}$$

For the reasons concerning the choice of r, see Capursi and Librizzi (2007). In particular, the transformation (2.2) is obtained as a particular case of the complement to the unity of a relative index of dissimilarity between the ordinal empirical distribution of the judgments and the ordinal distribution 'excellent', namely the utmost agreement on the best judgment (Leti, 1983). So (2.2) gives a quantitative variable for each item and the statistical unit is the single teaching course. Moreover, the indicator allows to discriminate the items with the same median. Among all the items with the median *more yes than no*, the highest value (0.87) corresponds to items F3 and F5, for which 62.6% and 64.5%, respectively, of the opinions are positive.

% observations in each category							
Item	1	2	3	4	n. observed	median	$IS_{0.5}$
B3	5.7	15.4	39.4	39.6	8373	3	0.76
B4	8.0	17.4	35.1	39.5	8395	3	0.73
B8	8.8	18.4	43.5	29.3	8435	3	0.69
B10	13.1	19.3	40.6	27.0	8435	3	0.65
B11	10.8	23.9	43.7	21.6	8398	3	0.64
C2	9.4	18.0	40.0	32.6	8446	3	0.70
D1	14.0	24.4	42.1	19.5	8432	3	0.60
D2	22.6	34.9	33.1	9.3	8404	2	0.46
D3	12.8	20.2	37.5	29.5	8335	3	0.65
E1	12.2	20.6	39.9	27.2	8448	3	0.65
F2	5.4	8.9	29.3	56.4	7489	4	0.81
F3	2.6	6.3	28.4	62.6	8387	4	0.87
F4	3.5	7.4	38.1	51.0	8122	4	0.83
F5	2.6	5.8	27.1	64.5	8380	4	0.87
F6	9.4	15.5	37.0	38.1	8401	4	0.72
F7	9.3	14.3	35.3	41.1	8393	4	0.74

Table 2.4: Answers to items: 1 = *definitely no*, 2 = *more no than yes*, 3 = *more yes than no*, 4 = *definitely yes*

Chapter 3

Aggregation analysis: Relative Importance Metric

In this chapter we describe the methodology and results concerning the first phase of the work. In particular we want to investigate what are the items that explain the satisfaction. Because our study is observational with correlated variable (see Appendix B), we make use of Relative Importance Metric (RIM) in linear regression to estimate the weight for each item in explanation of satisfaction. In this phase of work the data are transformed by means 2.2 (Chapter 2, Section 2.5), so that the variable entering in the regression model, i.e. the items, are quantitative.

Moreover, under the assumption that the performance of students' careers can affect the expressed opinions, we use a performance indicator. This indicator, described in the next section, is built on the basis of the information obtained through the questionnaire filled out by students. This is intended to verify whether the drivers of the quality of teaching are different depending on the performance.

We introduce the Indicator of Student Performance in Section 3.1. Section 3.2 and 3.3 present the relative importance metric PMVD (Proportional Marginal Variance Decomposition) and a statistical test to compare PMVD metric for two groups of students (*bad* and *good* students). The results of the application to teaching evaluation data (Chapter 2) are shown in Section 3.4.

3.1 The indicator of Student Performance

Since the questionnaire is anonymous, the only way to know the performance of students is using the information included in the questionnaire. Such information declared by students, relate to credits gathered, age, sex, school of provenance ... By combining the variables age and acquired credits we build a performance indicator (ISP) (Librizzi, 2008) shown below:

$$ISP = (A - 19) * 0.8 - C/60. \tag{3.1}$$

A indicates the student age declared by him/her in the day when the questionnaire was filled out, C indicates the variable credits (the credits he/she says to have gathered). C are divided by 60 to express them in terms of 'fruitful years', given that students should acquire 60 credits per year. We subtract 19 from A, since it is the standard age for students to enter into the Italian university system. Therefore, the result is the number of years spent in university studies (assuming that students enter into the university system at the age of 19 exactly). This number is multiplied by 0.8, to adjust it to the standard of students performance, since students with an excellent career are very rarely observed. This is equivalent to assume that a student

reaches, on average, 48 credits per year.

Indicator 3.1 can take negative values and it has not a theoretical maximum, since there is no theoretical maximum for student age. To sort out this drawback, ISP is standardized in the following way:

$$ISP^* = 1 - \frac{ISP + k}{max(ISP + k)}, \tag{3.2}$$

where $k = -min(min(ISP), 0)$. In this way the second addend of 3.2 gets values between 0 and 1. For a straightforward interpretation we consider 1 minus the fraction. So, in according to our data ISP^* is equal to 0 when a student is 28 years old and he has just acquired 30 UEC; ISP^* is equal to 1 when ISP is equal to its maximum that is obtained crossing age 20 with the higher number of observed credits. ISP^* allows,to classify students in *bad* and *good* relating to their performance.

3.2 Relative importance metrics

Weighting techniques based on a multiple regression model are widely used because of the numerous advantages that such techniques involve, like the possibility to determine the weight of the single simple indicators (Nardo *et al.*, 2005). When regressors are uncorrelated each covariate contribution is just the R^2 from univariate regression, and all univariate R^2-values add up to the full model R^2. But, when data come from observational studies, the covariates are usually correlated and such techniques are not appropriate because it is not simple to break down R^2 into components from the individual regressors. Let consider the linear regression model

$$Y = \beta_0 + X_1\beta_1 + ... + X_n\beta_n + \epsilon \tag{3.3}$$

where random variables X_j, $j = 1, ..., n$, denote n regressor variables and ϵ denotes an error term with expectation 0 and variance σ^2. This model implies $E(Y|X_1, ..., X_p) = \beta_0 + X_1\beta_1 + ... + X_p\beta_p$ and $var(Y|X_1, ..., X_p) = var(\epsilon|X_1, ..., X_p) = \sigma^2$. The marginal variance model is

$$var(Y) = \sum_{j=1}^{n} \beta_j^2 v_j + \sum_{j=1}^{n-1} \sum_{j+1}^{n} \beta_j \beta_k \sqrt{v_j v_k} \rho_{jk} + \sigma^2. \tag{3.4}$$

The regression variances are denoted as v_j, $j = 1, ..., p$, the inter-regressor correlations as ρ_{jk}. If X's are uncorrelated, the explained variance can be split into the contribution $\beta_j^2 v_j$ ($v_j = var(X_j)$), can be consistlently estimated using the unique sum of squares for each regressor. If X's are correlated, it is not possible to decompose $var(Y)$ in the usual way.

The difficulty in decomposing R^2 for regression model with correlated regressors lies in the fact that each order of regressors yields a different decomposition of the sum of squares (Achen, 1982). Generally the regressors enter into the model in the order they are listed.

In 1982 Achen has introduced a distinction between "dispersion importance", i.e., importance relating to the amount of explained variance, "level importance", i.e., importance of each regressor for the response's mean, or "theoretical importance" i.e., change in the response for a given change in the regressor. Some scholars have proposed analytical procedures able to underline the relative importance of each variable within a regressive model (Firth, 1998). Nevertheless, these various approaches have not found unanimous agreement because of the different results reached in presence of correlation among the regressors. Moreover, if we consider a regression model we can observe that regressors are significant, but among these we cannot determine a ranking of the regressors or a quantification of the rela-

tive importance of each regressor for the response.
Approach to this issue are proposed in literature by means of relative importance metrics for the R^2 decomposition (Feldman, 2006, 2007; Lindeman, 1980).

In literature the more used metrics are LMG (Lindeman Merenda Gold) and PMVD (Proportional Marginal Variance Decomposition). Both metrics (Lindeman, 1980; Feldman, 2006, 2007) decompose R^2 into non-negative contributions that automatically sum to the total R^2.

The approach taken by the metrics LMG and PMVD is based on sequential $R^2 s$. It takes into account the dependence on orderings by averaging over orderings (Kruskal, 1987a,b), either using unweighted averages (LMG) or weighted averages with data-dependent weights (PMVD).
The following criteria for decomposition of the model R^2 are considered useful in the literature, though seldom listed explicitly:

a) Proper decomposition: the model variance is to be decomposed into shares, and the sum of all shares has to be the model variance.

b) Non-negativity: all shares have to be non-negative.

c) Exclusion: the share allocated to a regressor X_j with $\beta_j = 0$ should be 0.

d) Inclusion: a regressor X_j with $\beta_j = 0$ should receive a non zero share.

Feldmann (Feldman, 2006) critized that LMG violates the exclusion criterion (for which the share allocated to a regressor X_j with $\beta_j = 0$ should be 0) and designed PMVD specifically for satisfying this criterion. If a causal interpretation of the variance allocations is intended, LMG's equalizing behavior must be seen as a natural result of model uncertainty and LMG is

to be preferred (Grömping, 2007). In our study we prefer PMVD metric for two reason. First of all our aim isn't to find the causal link between items taking into account the correlation structure between items; secondly we consider exclusion an indispensable criterion in the application. For describing the metric PMVD, we introduce the following notation. In linear regression the coefficients β_k, $k = 0, ..., p$ are estimated by minimizing the sum of squared unexplained parts. Denoting $\hat{y}_i$ the fitted values and considering a set $\mathcal{S}$ of p regressors, R^2 is given by the ratio between regression deviance and total deviance:

$$R^2(\mathcal{S}) = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}. \tag{3.5}$$

R^2 measures the proportion of variation in y that is explained by the p regressors in the model.
The sequentially added explained variance, obtained when we add the regressors with indices in $\mathcal{M}$ to a model that already contains the regressors with indices in $\mathcal{S}$ is gives as

$$seqR^2(\mathcal{M}|\mathcal{S}) = R^2(\mathcal{M} \vee \mathcal{S}) - R^2(\mathcal{S}). \tag{3.6}$$

The order of the regressors in any model is a permutation of the regressors $x_1, ..., x_p$. It is denoted by $r = (r_1, ..., r_p)$. Let $S_k(r)$ the set of regressors entered into the model before regressor x_k, according to the order r, then the portion of R^2 allocated to regressor x_k in the order r can be written as

$$seqR^2(\{x_k\}|S_k(r)) = R^2(\{x_k\} \vee S_k(r)) - R^2(S_k(r)). \tag{3.7}$$

As said, PMVD can be seen as an average over orderings as well, with data-dependent weights accordind to the r-th order:

$$PMVD_k = \frac{1}{p!} \sum_{r=1}^{p!} w(r) seqR^2(\{x_k\}|r), \tag{3.8}$$

where $w(r)$ denotes the data-dependent weights. In this case, if the coefficients of the regressors are not zero, the permutation r has a weight proportional to

$$L(r) = \prod_{i=1}^{p-1} seqR^2(\{x_{r_{i+1}}, ..., x_{r_p}\}|\{x_{r_1}, ..., x_{r_i}\})^{-1} \tag{3.9}$$

and

$$w(r) = L(r)/\sum_r L(r) \tag{3.10}$$

is the probability associated to the order r, where summation in the denominator is over all possible permutations r. In other words, PMVD weights are obtained through a weighted mean of increases R^2 over all possible entry orders. Feldman's proposal (Feldman, 2007) gives a weigth proportional to the R^2 explained by each regressor. This implies that the distribution of relative importance measures is concentrated on few regressors with high predictive power.

3.3 Are good and bad students significantly different?

To answer to this question, it is necessary to construct a statistical test to compare, for every item $k = 1, ..., K$, the weights obtained with PMVD met-

ric for two grous. Because we have not standard error of PMVD, we utilize bootstrap procedure to construct an empirical sampling distribution and to assess the reliability of relative importance measures (Efron and Tibshrani, 1993). To build the statistical test, for two groups, we resample 500 times the values PMVD for every item, obtaining two matrices M_1 and M_2 of dimension $500xK$. Then, relating these matrices, we obtain the ratio matrix R with generic element r_{ik}, where $i = 1, ..., 500$, 500 is the sample dimension and $k = 1, ..., K$ indicates the item. The joint distribution of the K distributions r_k is a multinormal distribution. From R matrix we determine the variance and covariance matrix bootstrap $V^{\star}(\hat{R})$ of dimension KxK. The statistical test is the following (Dobson, 1983):

$$\hat{r}^T V^{\star}(\hat{R})^{-1} \hat{r}, \tag{3.11}$$

with a χ^2_K distribution, where $\hat{r}$ is the ratio vector of observed weights PMVD between two groups.

3.4 Results

3.4.1 Indicator of Students Performance results

In this section we present some considerations on indicator (3.2), justifying the classification of students in relation to their performance. The graphic representation of conditional distribution of ISP^* given age (Figure 3.1) highlights an increasing monotonous trend of median level of non regularity to the growth of age. So, the variability of ISP^* is explained by age.

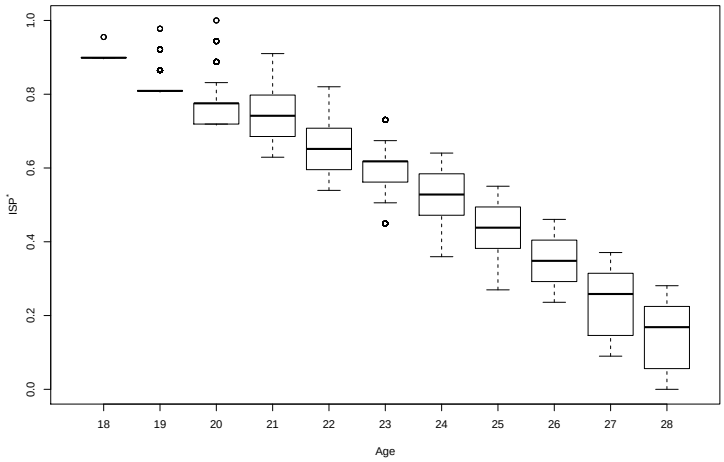

Figure 3.1: Boxplot of conditional distribution of ISP^* given *age*.

Other considerations on indicator (3.2) can be drawn from Figure 3.2

- in this graphic we can observe the level curves of ISP^*;
- the lowest values of the indicator are obtained for the students who have a very bad career;
- the indicator increases for decreasing values of age and/or increasing values of credits (UEC);
- we can observe that in the top right side of the graphic there are not any observed values, because it is not possible that a student is ahead of schedule;
- dots size highlights a very high frequency of 19 students years old who acquired 30 UEC. So we can consider that values between 0.7

and 0.8 correspond to a standard career. For example, this interval comprehends students who achieve a first degree (180 UEC) at 22 or 23 years old.

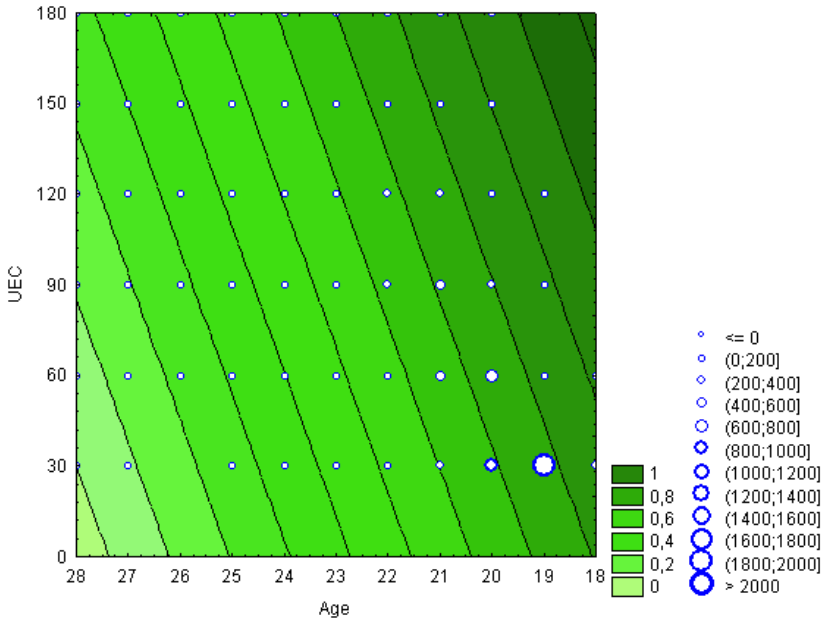

Figure 3.2: Level curves of ISP^* as a function of *age* and *UEC* with frequency classes of students.

In Table 3.1 we can observe the frequency of distribution for classes of values of ISP^*. The three classes of values greater than 0.7 represent positive results. In particular more than half (about 70%) of students have an excellent or standard career.

Classes of values	Frequency	% frequency
0-0.1	33	0.39
0.1-0.2	36	0.49
0.2-0.3	100	1.18
0.3-0.4	140	1.65
0.4-0.5	266	3.13
0.5-0.6	742	8.74
0.6-0.7	2400	28.28
0.7-0.8	3982	46.91
0.8-0.9	755	8.89
0.9-1	34	0.40

Table 3.1: Distribution of students for classes of values of ISP^*.

These empirical considerations lead us to define the following dichotomous variable:

$$P = \begin{cases} 0 & \text{if } ISP^* \leq 0.7 \\ 1 & \text{if } ISP^* > 0.7. \end{cases}$$

P takes value 0 if the student time lag is greater than the standard one, i.e. if he has a bad career performance. On the other hand, when the time lag indicator is greater than 0.7 ($P = 1$), we consider the student has a good career performance.

3.4.2 Relative importance metric results

Students satisfaction depends on several aspects of the teaching activities, but not all with the same importance. We are interested in identifying which

items are the drivers of quality of teaching in the students opinion (Campostrini *et al.*, 2006), as in Capursi et al. (V. Capursi, 2008), and above all, in highlighting possible differences between *good* and *bad* students. The complexity of the concept that we want to measure makes necessary to pay attention to analysis of data. In fact, evaluation items of the questionnaire are highly correlated, so it is difficult to identify those that greatly influence the global satisfaction. We use relative importance metric to try to obtain weights that more explain students satisfaction. Before relative imporantance analysis, original ordinal data were aggregated by teaching course by means 2.2.

PMVD metric

To find the relative importance of such items, we use PMVD method on the basis of a linear model in which the indicator (2.2) for item C2, is regressed on indicator of questionnaires items. We consider item C2 because we think that C2 can express the general perception of teaching quality from students. Initially, we consider a model in which ISP^* variable is present. Because of high correlation of item covariates, the effect of this variable is non relevant. For this reason we consider two separated models for the two groups of students (*bad* and *good*):

$$\begin{aligned} I_{C2i} =& \beta_{0i} + \beta_{1i} IS_{B3i} + \beta_{2i} IS_{B4i} + \beta_{3i} IS_{B8i} + \beta_{4i} IS_{B10i} + \beta_{5i} IS_{B11i} + \\ &+ \beta_{6i} IS_{D1i} + \beta_{7i} IS_{D2i} + \beta_{8i} IS_{D3i} + \beta_{9i} IS_{E1i} + \beta_{10i} IS_{F2i} + \\ &+ \beta_{11i} IS_{F3i} + \beta_{12i} IS_{F4i} + \beta_{13i} IS_{F5i} + \beta_{14i} IS_{F6i} + \beta_{15i} IS_{F7i} + \epsilon_i, \end{aligned} \tag{3.12}$$

the first one ($i = 0$) for *bad* students and the second one ($i = 1$) for *good* students. Results are shown in Table 3.3, where PMVD weights are scaled

so that they sum to 1 to make interpretation easier. First of all, we can observe that many items have weigth zero or almost zero. In particular, for both groups, items refer to organization of teaching (B11, D1, D2, D3, E1) have small weigths. R^2 is equal to 0.78 for the first model and 0.88 for the second one.

Observing the weights, the items that explain more the students satisfaction, in term of relative importance, are B3 and F7 for *bad* students; the more importance items for *good* students B3, F6 and F7. So there are some differences. For *good* teacher motivation (F6) is more important than *bad*. We observe a weigth of 0.282 for *good* and 0.075 for *bad*. Moreover, the first group of students give a great importance than the second group to the clarity of teaching (F7) (0.647 vs 0.545). *Bad* students give small weigths to items F2, F3, F4, F5 (that refer to the teaching. For *good* unique important items in section F are F6 and F7. The *good* students are very demanding than *good* respect to clear explanation of formative objective of the teaching (B3) (0.132 vs 0.165). It seems that, somehw, the career performance, can be an element of discrimination to evaluate the teaching quality.

	PMVD	
Items	P = 0	P = 1
B3	0.165	0.132
B4	0.000	0.001
B8	0.019	0.019
B10	0.026	0.009
B11	0.008	0.001
D1	0.000	0.002
D2	0.016	0.000
D3	0.003	0.001
E1	0.004	0.003
F2	0.002	0.005
F3	0.022	0.000
F4	0.031	0.000
F5	0.021	0.000
F6	0.035	0.282
F7	0.647	0.545

Table 3.2: PMVD weights of teaching quality items.

Boostrap results

Summary statistics from bootstrap procedure are presented in Table 3.3. Mean bootstrap values overlap with the observed PMVD weigths (Table 3.2). In fact, for *bad* students, items B3 and F7 are the only items with higher weight. For *good* students, items B3, F6 and F7 have mean values greater than other mean value items. We observe OLS (Ordinary Least Square) analysis results (Table 3.4). For *bad*, items B3, B10, B11, D2, F3,

F4, F5 and F7 have significant coefficients. If we consider mean value in Table 3.3, these items (except for B3 and F7) have low weigths. For *good* students items B3, B8, B10, F2, F6 and F7 have significant coefficients. If we consider mean value in Table 3.3 these items (except for B3 F6 and F7) have low weigths. So, through PMVD metric we obtain different results in terms of importance in the explanation of overall satisfaction.

Considering that the excess kurtosis of the normal distribution is zero, the Bera Jarque p.values are based on the Bera Jarque test statistic and represents the confidence level in rejecting the hypothesis of asset return distribution normality based the sample values for the skew and kutorsis of the distribution. This test statistic is distributed χ^2. According to this test, the hypothesis that residuals are normal cannot be accepted only for any item.

Figure 3.3 shows the univariate distribution of PMVD component shares for all items for two student groups. It is evident that there are two types of distributions: highly skewed distributions almost exponential in nature such as observed for items F7 for both groups; symmetric kurtotic distribution such as observed for F6 in *good* students. In particular items with a low weight PMVD are approximately exponential, items with a high explanatory power in terms of relative importance have skew and kurtosis values lower that others. Now, we concentrate just on items B3, F6 and F7. We note that for item B3 and F7 there is the overlapping between the two curve. For item F6 there is a difference between two groups. In particular, for *good* students F6 has higher frequencies for high values of PMVD than *bad* students. Moreover we can oserve the non-overlapping between the two curves.

The null hypothesis of statistical test (3.11) is:

$$H_0 : \beta_{k0} = \beta_{k1},$$

where β_{k0} and β_{k1} are the coefficients of model (3.12) with $k = 1, ..., 15$ for $i = 0$ (*bad* students) and $i = 1$ (*good* students). Considering this statistical test with χ^2_{15} distribution, for $\alpha = 0.05$ we can reject the null hypothesis of equality of weights between two groups, that can be considered, in terms of relative importance, statistically different. It seems that this overall difference can be due to item F6 for which *good* students give more importance than bad.

	bad						good					
Item	Mean Value	Std. Dev.	Skew	Excess Kurtosis	BJ-stat	BJ-p.value	Mean Value	Std. Dev.	Skew	Excess Kurtosis	BJ-stat	BJ-p.value
B3	0.183	0.106	0.724	0.181	44.352	0.000	0.135	0.071	0.725	0.395	47.062	0.000
B4	0.004	0.006	2.608	7.548	1753.599	0.000	0.001	0.002	1.605	3.114	416.621	0.000
B8	0.028	0.034	2.321	7.019	1475.178	0.000	0.021	0.018	1.427	2.551	305.128	0.000
B10	0.026	0.020	1.298	1.956	220.135	0.000	0.009	0.007	1.572	4.970	720.448	0.000
B11	0.011	0.015	2.208	5.639	1068.800	0.000	0.002	0.002	3.456	17.655	7489.009	0.000
D1	0.004	0.005	2.330	6.919	1449.699	0.000	0.004	0.005	2.924	14.844	5303.313	0.000
D2	0.020	0.023	1.571	2.315	317.217	0.000	0.001	0.002	3.154	11.427	3548.959	0.000
D3	0.004	0.004	2.119	6.720	1314.843	0.000	0.001	0.001	1.454	2.082	266.528	0.000
E1	0.004	0.004	1.743	3.846	561.249	0.000	0.004	0.004	1.586	2.568	346.924	0.000
F2	0.003	0.003	2.049	5.625	1009.010	0.000	0.006	0.006	1.787	4.382	666.083	0.000
F3	0.025	0.021	1.475	2.568	318.619	0.000	0.001	0.003	4.875	32.219	23606.780	0.000
F4	0.034	0.039	1.962	4.440	731.393	0.000	0.002	0.005	5.016	28.306	18789.296	0.000
F5	0.024	0.013	0.743	0.646	54.661	0.000	0.003	0.005	4.244	23.460	12967.424	0.000
F6	0.069	0.186	1.794	3.238	486.642	0.000	0.267	0.112	0.435	-0.277	16.840	0.000
F7	0.561	0.126	-0.455	0.250	18.574	0.000	0.542	0.122	0.289	-0.204	7.836	0.020

Table 3.3: Boostrap statistics

In fact, considering the statistical test (3.11) with a χ^2_{15} distribution, for $\alpha = 0.05$ we can reject the null hypothesis of equality of weights between two groups, that can be considered, in terms of relative importance, statistically different.

	bad				good			
Par.	Beta	Std. Err.	t-stat	p-val	Beta	Std. Err.	t-stat	p-val
Interc.	-0.108	0.059	-1.831	0.068	-0.169	0.052	-3.264	0.001
B3	0.273	0.073	3.741	0.000	0.247	0.054	4.584	0.000
B4	0.016	0.052	0.304	0.761	-0.039	0.041	-0.965	0.335
B8	0.111	0.045	2.470	0.014	0.099	0.034	2.843	0.004
B10	0.100	0.038	2.600	0.009	0.077	0.028	2.685	0.008
B11	0.085	0.036	2.336	0.020	0.031	0.027	1.166	0.245
D1	-0.019	0.047	-0.404	0.686	0.050	0.041	1.198	0.232
D2	0.105	0.045	2.325	0.021	0.003	0.039	0.073	0.942
D3	-0.056	0.040	-1.414	0.158	-0.041	0.030	-1.356	0.176
E1	-0.051	0.035	-1.452	0.148	0.047	0.031	1.493	0.137
F2	-0.067	0.056	-1.192	0.234	0.092	0.041	2.248	0.025
F3	0.257	0.071	3.614	0.000	-0.000	0.058	-0-003	0.997
F4	0.215	0.066	3.242	0.001	-0.028	0.059	-0.482	0.630
F5	-0.374	0.087	-4.278	0.000	0.008	0.072	0.114	0.909
F6	0.094	0.061	1.532	0.127	0.258	0.054	4.809	0.000
F7	0.448	0.065	6.901	0.000	0.392	0.043	9.065	0.000

Table 3.4: OLS analysis

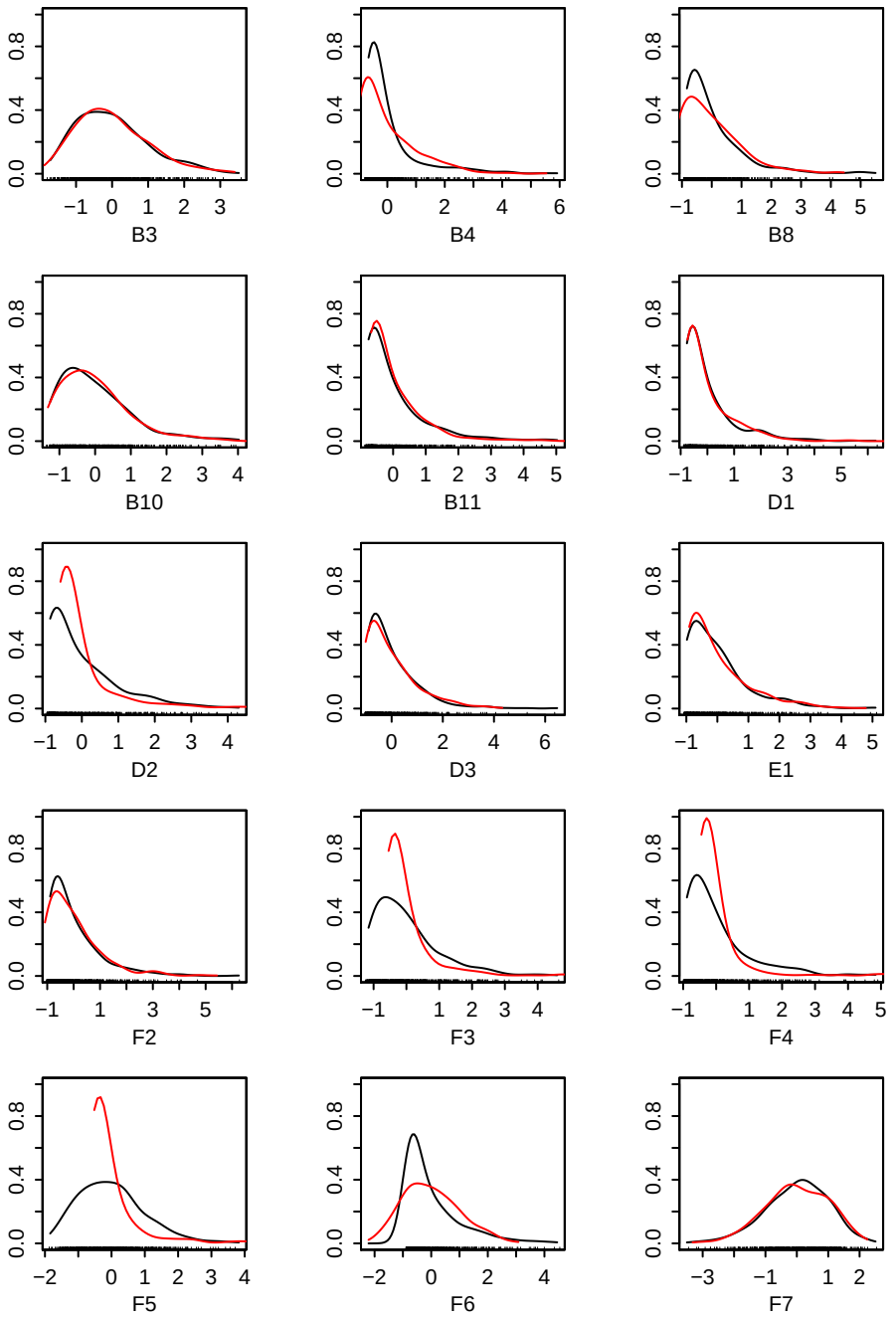

Figure 3.3: PMVD component boostrap distribution for bad (——) and good(——) students.

Chapter 4

Multilevel IRT Model

4.1 Why Multilevel IRT Model

In Chapter 3 we introduce relative importance metrics to determine the drivers of teaching quality. In this Chapter we want to use other models in order to take into consideration the complexity of Student Evaluation Teaching data. One of such models are Item Response Theory (IRT) models (P. de Boeck, 2004)

Our data are organized as a hierarchical structure (students in course, courses in courses degree). In this case, it is reasonable to expect that latent variable levels in the lower part of the hierarchy (students) are correlated to a greater extent with those belonging to different higher level units. In other words, it may be supposed that students within courses are not independent: students are evaluating the same course, they shared for a term the same lecturer, the same class environment and the same group rules: probably they have shared their opinion of the course during the term, affecting the final opinion on each other (Snijders and Bosker, 1999; Goldstein, 2002).

Within the framework of multilevel (or random effects or generalized linear mixed) models, an item response model (Rasch, 1980) is embedded in a hierarchical model. This framework is characterized by the treatment of person ability parameters as random parameters in a IRT model.
The importance of hierarchical structure has been well known in statistics for a long time. Methodological developments try to include hierarchies in analysis. In particular two approaches were considered and used. In the first approach it is estimated a single regression model for individual data, ignoring the presence of groups. In the second, multiple regression models are estimated, one for each group. So considering the first approach, we estimate a single Rasch model for all students, without distinguishing between the various courses, in the second we estimate a series of Rasch models, one for each course. These two solutions are very simple to apply but do not take into account in an appropriate way the structure of the aggregate data.
Using multilevel models we treat the data taking into account their hierarchical structure and the dependence of responses through the random effects. The Chapter is structured as follows: Section 4.2 describes the literature for the application of multilevel IRT model, the third section (Section 4.5) refers to the different procedures with which to develop a multilevel model. In particular, the two-step analysis is described, highlighting the possible issues; Section 4.3 is devoted to a brief description of IRT models for binary and ordinal data, introducing and presenting the main features of the Rasch model; in the fifth section (Section 4.4) we introduce the interpretation of the parameters of the IRT models in the context of the evaluation of teaching. In Section 4.6 we present a multilevel model as framework for IRT model both for binary and ordinal data. In particular we describe the

algebraic equivalence between multilevel ordinal model and ordinal IRT model. Moreover, in Section 4.7 we show via simulation study the algebraic equivalence between the two models.

4.2 Literature background

Literature contains various applications for binary multilevel IRT models. Verhelst and Eggen (1989) and Zwinderman (1997, 1991) considered the combination of an IRT model with a structural linear regression model.Raudenbush and Sampson (1999) discussed a multilevel model that can be seen as a Rasch model embedded within a hierarchical structure, where the first level of the multilevel model describes the relation between the observed item scores and the ability parameters. Kamata (2001) introduced a multilevel formulation of the Rasch model using HLM software. Fox and Glas (1991) and Pastor (2003) explored and illustrated the use of Kamata's three level IRT model in educational and psychological measurement and research. As concerns polytomous data, Maier (2001) uses a hierarchical Partial Credit Model (PCM), with covariates at the level of individuals, to determine whether gender differences existed in the student's mood in a mathematics classroom. Fox (2001) estimates multilevel IRT models with latent dependent and independent variables and dichotomous and polytomous items, in order to assess the school effectiveness. Adam *et al.* (1997) and Patz and Junker (1999) discussed models that can handle both dichotomous and polytomous item responses along with a latent variable as outcome in a regression model.

4.3 Theoretical background: the Rasch model

The Rasch Model (RM) was the simplest model among the IRT models. It was first proposed in the 60s to evaluate ability tests (Rasch, 1960). The RM is a latent structure model by means of which it is possible to derive continuous measures on an interval scale from total scores obtained by a set of subjects on a set of items. This situation is common in social sciences, as stated, for example, by Molenaar in a fundamental book on Rasch Model (Fischer and Molenaar, 1995): "It is easy to find examples of observable human behaviour indicating that a person has more or less of such a general property, but the concept has a surplus value, in the sense that no specific manifest behaviour fully covers it. This is the reason why such properties are called *latent traits*". The fundamental assumption of the Rasch model is that the answer each subject gives to each item depends on two parameters: one is the *person* parameter and represent a subject measure (θ_j), the other is the *item parameter* that is the item measure (π_i). Then the response probability of each subject to each item is a function of person and item parameters. It is possible to compare these two parameters because they belong to the same continuum. Their interaction is expressed by the difference $\theta_j - \pi_i$ ($j = 1, ..., J$, $i = 1, ..., I$). In a deterministic sense a positive difference means that the subject's abilities are superior to the item's difficulty and therefore we can be sure that an exact response will always have been given. From a probabilistic perspective, such as that of the RM, this is not true since a subject who is intrinsically capable of giving a right answer ($\theta_j > \pi_i$) may instead, given a wrong response. Likewise, it is possible that a subject lacking in ability can accidentally give a right answer.

The more simple Rasch Model is the dichotomous one. In this case, the probability of a correct answer $Y_{ij} = 1$ by the subject j of ability θ_j when

answer to the item i of difficulty π_i is:

$$P(Y_{ij} = 1) = p_{ij} = \frac{exp[\theta_j - \pi_i]}{1 + exp[\theta_j - \pi_i]} = \frac{1}{1 + exp[-(\theta_j - \pi_i)]} \tag{4.1}$$

In the dichotomous model data are collected in the *raw score matrix*, with J rows (one for each subject) and I columns (one for each item), whose values are equal to 0 or 1. The sum of each row $r_j = \sum_{i=1}^{I} y_{ij}$ represents the total score of the subject j for all the items; the sum of each column $s_i = \sum_{j=1}^{J} y_{ij}$ represents the score given by all subjects to the item i. These scores are given according to a metric that, being nonlinear, produces some conceptual distorsion when we compare the row and column totals. So, it is necessary to change these scores according to a metric that is founded on the conceptual distances between subjects and items (Wright and Masters, 1982). The transformation takes place through the logit:

$$log \frac{p_{ij}}{1 - p_{ij}} \tag{4.2}$$

Some assumptions are fundamental in all family of Rasch models parameters. The first is that the items measure only one latent feature (*unidimensionality*. Another important characteristic is that the answers to an item are independent of the answers given to other items. As far as the parameters are concerned, for which no assumptions are made, by applying the logits previously described, θ_j and π_i can be expresses according to a common measurement unit on the same continuum (*parameters linearity*); the estimation of person ability (θ_j) is free from sampling distribution of the items attempted; the estimation of the item difficulty (π_i) is free from the sampling distribution of the sample employed (*paramters separability*); and

the row and column totals on the row score matrix are sufficient statistics for the estimation of θ_j and π_i. A fully examination of these assumptions is beyond the scope of this paper. For a detailed discussion see (Fischer and Molenaar, 1995).

4.3.1 Polytomous IRT model

The Rasch dichotomous model has been extended to the case of more than two ordered categories. In this model it is introduced the assumption that between each category and the next there is a threshold that qualifies the item's position and characterizes the π_i as a function of the difficulty presented by every answer category. Thus the answer to every category m depends on the value π_{im} that is the category difficulty for category m and item i. In particular $\pi_{im} = \pi_i - c_m$ is the difference between the location parameter for item i and the category threshold for category m.
Different politomous models have proposed:

- the Rating Scale Model (RSM) (Andrich, 1978). A fundamental condition of the RSM is the equality of the threshold values for all the items: even if the distance between a threshold and another one can differ, the pattern of these distances is constant for all the items;

- the Partial Credit Model (PCM)(Masters, 1982). In this model the difficulty levels differ item by item and the subject receives a partial credit (score for each item) equivalent to the relative level of difficulty of the completed performance. The thresholds can differ freely in the same item or from one item to another.

The IRT model considered in our analysis is the PCM. Denote with p_{mij} the probability of person j to respond with category m to item i, assuming

for item i there are M ordered response categories ($m = 1, ..., M$). Then this probability is:

$$P(Y_{ij} = m) = \frac{exp[\theta_j - \pi_{im}]}{1 + exp[\theta_j - \pi_{im}]} \tag{4.3}$$

This function depends from the person parameters θ_j, the item parameter π_i and the threshold parameters δ_m that measure additional difficulty to endorse the m-th response category. These parameters represent the cut-off between one response category and the following

4.4 Rasch Model and teaching quality

In order to apply the Rasch model for measuring the quality of teaching, we need to find a correct interpretation of the parameters taking into account the context of teaching evaluation. Two different factors are often confused but are very different. Let us start with an example: we want to measure the quality of ice cream in Italy and in Germany. Probably Italian people about are less satisfied than Germany people. The quality of ice cream is the same. Satisfaction differs because of cultural differences, traditions. So we can consider quality as the attribute factor and satisfaction as the person factor and together these factors determine the result of the single answer in the questionnaire. We can define now the correspondence between original application (in psychometrics) and service quality application of the Rasch model. The factor related to the persons, that in original application was the ability, become now the satisfaction. The factor related to the items that was the difficulty, in quality service becomes the quality (Figure 4.1).

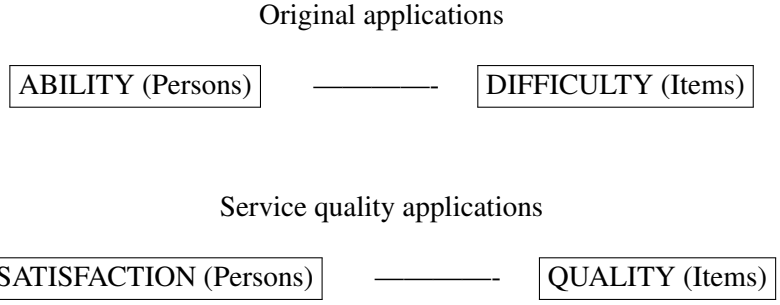

Figure 4.1: Correspondence between original and service quality applications

In the original context the scale of the π_i parameters is interpreted in the following way: the smallest values of the π_i parameters are associated to easy item (so the subjects have a high probability of exceeding the item's difficulty); while the highest values are associated to the more difficult items (the probability of overcoming the item's difficulty is lower). On the other hand, in quality context, the scale has to be read in the opposite way (Figure 4.2: the smallest values of the π_i parameters identify the items with good quality (because the subject satisfaction probabilities are high), while the highest values of the item parameters correspond to items with bad quality (lower subject satisfaction probabilities).

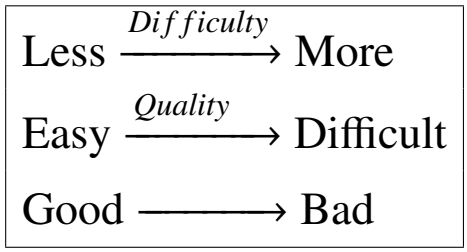

Figure 4.2: Item scale

For the scale of the parameters θ_j the interpretation is the same in both cases (Figure 4.3): the smallest values of the parameter, which identifies subjects of low ability, now identifies subjects with low levels of satisfaction, and the greatest values, which previously corresponded to subjects with a high degree of ability, now correspond to subjects with a high level of satisfaction.
Henceforth π_i will be named *item quality* and θ_j *person's satisfaction*.

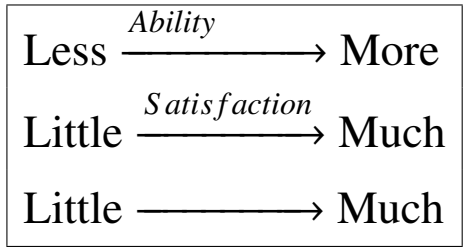

Figure 4.3: Person scale

4.5 Two step analysis vs one step analysis

A two step analysis using Item Response Theory (IRT) models is a common practice, especially in investigating effects of student characteristics on student abilities. In such a two-step analysis, student abilities are estimated via a standard IRT model as the first step. Then, in the second step, ability estimates are used as an outcome variable, and student characteristic variables are used as predictors in a simple linear model, such as multiple regression and analysis of covariance. For example in Pagani and Zanarotti (2010), students' satisfaction, measured through person location parameters (PLP) obtained with the Partial Credit Rasch model, was used to assess the quality of teaching service in a set of university courses. Firstly, the authors

use the PLP to obtain a measure of the level of satisfaction, then they introduce this measure as a dependent variable in a multilevel model to detect individual and environmental determinants of the level global satisfaction. Lucadamo (2010) the purpose is to evaluate the Customer Satisfaction of the patients of an hospital. The quantification of the response is made by the use of the Rasch analysis, and in the second step of the work, he tries to verify if the patient satisfaction can be influenced by socio-economic factors using a multilevel model. Rampichini and A. Petrucci (2004) present a methodology for the analysis of student ratings of university courses. First they discuss simple descriptive measures that take into account the ordinal nature of the ratings; then they present net measures which account for the characteristics of the students. These measures are obtained through multilevel modelling.

These two step analyses may not provide accurate results, because of at least two reasons. First, the standard error ability estimates from an IRT model are heteroschedastic. Second, it is known that person parameter estimates from marginal maximum likelihood estimation are biased and inconsistent (Goldstein, 1980). Through a single analysis rather than two-step analysis, one can expect improved estimation of the effects of student characteristic variables on a latent trait, because these effects are estimated simultaneously with ability parameters. As a result, the heteroschedastic nature of the standar errors of ability parameters, is to take into account.

In the next Chapter we present two-step and one-step analysis results, in order to highlight some differences in outcome.

4.6 Multilevel Framework for IRT model

In this Section we present the multilevel framework for IRT model. In particular we demontrate the algebraic equivalence between multilevel and IRT models. This framework we allow us to carry out a one-step analysis. We pay more attention on ordinal IRT model. For greater comprehension of the equivalence, we present, initially the multilvel framework for the simplest IRT model: the Rasch Model (RM). Next, we present multilevel framework for an ordinal IRT model: Partial Credit Model (PCM). In particula we show the equivalence between a three-level model and the ordinal PCM. This provide estimates of group-level satisfaction as well as person-level satisfaction. Moreover we add in the model a person characteristic to try to quantify the varioatio of person-characteristic variable effectc across groups.

4.6.1 Multilevel Framework for bynary IRT model

We can show the equivalence between multilevel and Rasch Model ((Kamata, 2001)) expressing the 4.1 by means logit link function. This function can be used to model the log odds of the probability of a correct response:

$$\eta_{ij} = \theta_j - \pi_i = log\left(\frac{p_{ij}}{1 - p_{ij}}\right) \tag{4.4}$$

where η_{ij} is the log odds of obtaining a correct response to item i for a person j. For a set of I items, the items could be modeled assuming a Bernoulli distribution and the logit link function such that η_{ij} becomes:

$$\eta_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + ... + \beta_{k-1}X_{(k-1)ij} = \beta_{0j} + \sum_{q=1}^{k-1}\beta_{qj}X_{qij} \tag{4.5}$$

where X_{qij} is the q-th dummy variable for person j, with values -1 when $q = i$, and 0 when $q \neq i$ for item i. β_{0j} is an intercept term, and β_{qj} is a coefficient associated with X_{qij}, where $q = 1, ..., k-1$. For item i associated with the q-th dummy variable the equation 4.5 becomes

$$\eta_{ij} = \beta_{0j} - \beta_{qj}. \tag{4.6}$$

Note that no indicator variable is associated with the k-th item because it is assumed that $\beta_{kj} = 0$ to warrent the full rank of the design matrix. β_{0j} is an intercept term, and a value 1 is assigned to X_{0ij} for all observations. So, β_{0j} is considered to be an overall effect that is common to all items. On the other hand $\beta_{kj} = 0$ means that the effect of the k-th item, compared with the overall effect, is assumed to be zero. Then the probability that person j answers item i correctly is expressed as:

$$p_{ij} = \frac{1}{1 + exp(-\eta_{ij})}. \tag{4.7}$$

This is the level 1 or item level of the multilevel model, with items nested within students. At this level the β_s are constant across people. It should also be noted that the β_s are not the final parameters that are considered to be item difficulties. The item parameters are defined in the level-2 model, and they may be characterized as being constant across people.

The level-2 model is the person level model. Since β_{0j} is treated as a parameter common to all items in the level-1 model, it must be assumed in the level-2 models that β_{0j} is a random effect across people. In this way, a latent trait common to all items but variable across people can be modelled. Also while the level-1 model does not assume that β_{1j} through $\beta_{(k-1)j}$

are common across people, the level-2 models may model that item effects are constant across people by modelling the specifying $\beta_{qj}s$ as constants. Therefore, the level-2 model is:

$$\begin{cases} \beta_{0j} = \gamma_{00} + \mu_{0j} \\ \beta_{1j} = \gamma_{10} \\ \vdots \\ \beta_{(k-1)j} = \gamma_{(k-1)0} \end{cases} \tag{4.8}$$

where μ_{0j} is a random component of β_{0j} and is distributed as $N(0, \sigma_m^2 u)$. The level-1 model together with the level-2 models show that item parameters are fixed across people and vary across items, while a latent trait (person parameter) varies across people and fixed across items, because there are not random terms added into β_{1j} through $\beta_{(k-1)j}$. When level-1 and level-2 models are combined, the linear predictor model becomes $\eta_{ij} = \gamma_{00} + \mu_{0j} - \gamma_{q0}$ for person j and for a specific item i that is associated with q-th dummy variable. Then, combining 4.6 and 4.8 in 4.9, the probability that person j answers a specific item i correctly is expresses as

$$p_{ij} = \frac{1}{1 + exp\{-[\mu_{0j} - (\gamma_{q0} - \gamma_{00})]\}}, \tag{4.9}$$

where $i = q$. This has exactly the same form as the Rasch model in Equation 4.1, where $\theta_j = \mu_{0j}$, $\pi_i = \gamma_{q0} - \gamma_{00}$ for $q = i$.

4.6.2 Multilevel Framework for ordinal IRT model

In presence of ordinal data, very useful models are multilevel ordered logistic regression models, also called the multilevel ordered logit models or

the multilevel proportional odds models.

The multilevel ordered models can be formulated as threshold models. The real line is divided by threshold into M intervals, corresponding to the M ordered categories. The first threshold is δ_1. δ_1 defines the upper bound of the interval corresponding to observed outcome 1. Similarly, threshold δ_m defines the boundary between the intervals corresponding to observed outcomes $m-1$ and m. The latent response variable (teaching quality) is denoted by y^*_{ij} (for item i and student j) and the observed variable y_{ij} is related to y^*_{ij} by the threshold model defined as

$$y_{ij} = \begin{cases} 1 & \text{if } -\infty < y^*_{ij} \leq \delta_1 \\ 0 & \text{if } \delta_1 < y^*_{ij} \leq \delta_2 \\ \vdots & \\ M & \text{if } \delta_{M-1} < y^*_{ij} \leq +\infty \end{cases}$$

The Three-level model

Consider the latent response variable y^*_{ijg} for level-one units i (items), level-two units j (students) and level-three unit g (courses). The ordinal models can be written in terms of y^*_{ijg}

$$y^*_{ijg} = \eta_{ijg} + \epsilon_{ijg} \tag{4.10}$$

where

$$\eta_{ijg} = \beta_{0jg} + \sum_{q=1}^{k-1} \beta_{qjg} X_{qijg} \tag{4.11}$$

where X_{qijg} is the q-th dummy for student j in course g, with values -1 when $q = i$ and 0 when $q \neq i$. β_{0jg} is an intercept term and β_{qjg} is the coefficient

associated with X_{qijg}.
In absence of explanatory variables and random intercepts, the response variable y_{ijg} takes on the values of m with probability

$$p_{ijg}(m) = P(y_{ijg} = m), \tag{4.12}$$

for $m = 1, ..., M$. As ordinal response models often utilize cumulative comparisons of the ordinal outcome, they define the cumulative response probabilities for the M categories of the ordinal outcome y_{ijg} as

$$P_{ijg}(m) = P(y_{ijg} \leq m) = \sum_{k=1}^{m} p_{ijg} \qquad m = 1, ..., M \tag{4.13}$$

Note that this cumulative probability for the last category is 1. Therefore there are only $(M-1)$ cumulative probabilities to estimate. If the cumulative density function of ϵ_{ijg} is F, these cumulative probabilities are denoted by

$$P(y_{ij} \leq m) = F(\delta_m - \eta_{ij}) \qquad m = 1, ..., M-1, \tag{4.14}$$

Equivalently, we can write the model as a cumulative model

$$G[P_{ijg}(m)] = \delta_m - \theta_{ijg} \tag{4.15}$$

where $G = F^{-1}$ is the link function.
If ϵ_{ijg} follows the logistic distribution, this results in the multilevel ordered logistic regression model. Assuming the distribution of the error term ϵ^*_{ijg} of the latent response y^*_{ijg} to be logistic, the cumulative probability function of y_{ijg} will be written as

$$P_{ij}(m) = P(\epsilon_{ijg} \leq \delta_m - \eta_{ijg}) = \frac{exp(\delta_m - \theta_{ij})}{1 + exp(\delta_m - \theta_{ijg})}. \tag{4.16}$$

The idea of cumulative probabilities leads naturally to the cumulative logit model

$$log\left[\frac{P(y_{ij} \leq m)}{P(y_{ij} > m)}\right] = \delta_m - \eta_{ij} \tag{4.17}$$

Level-1 model

The level-1 model, which models variation of item responses within people, was used to model the log-odds of the probability of endorsing item i for student j in course g. The model is:

$$log\left[\frac{p(Y_{ijg} \leq m|x_{ijg}, \beta_{0jg})}{1 - p(Y_{ijg} \leq m|x_{ijg}, \beta_{0jg})}\right] = \delta_m - (\beta_{0jg} + \sum_{p=1}^{k-1} \beta_{qjg} x_{qijg}) \tag{4.18}$$

where δ_m is the threshold parameter for category $m = 1, ..., M - 1$. β_{0jg} represents the overall effect common to all items for person j.

Level-2 model

The second level or student-level of the model is used to model variation of students satisfaction level within course:

$$\begin{cases} \beta_{0jg} = \gamma_{00g} + \mu_{0jg} \\ \beta_{1jg} = \gamma_{10g} + \mu_{1jg} \\ \beta_{2jg} = \gamma_{20g} + \mu_{2jg} \\ \vdots \\ \beta_{pjg} = \gamma_{(k-1)0g} + \mu_{(k-1)jg} \end{cases} \tag{4.19}$$

The item effects ($\beta_{1jg}, ..., \beta_{(k-1)jg}$) are specified as random across students, so that the second level models the variation in β_{0jg} among students within courses, and the variation among students within courses for each item. It is assumed that the distribution of μ_{0jg} (the satisfaction estimates for students) is $N(0, \sigma^2_\mu)$. σ^2_μ represents the variation of satisfaction among students within courses.

Level-3 model

The third level of the model is used to model variation among courses in satisfaction using the parameters estimated for each course (γ_{00g}, γ_{10g},..., $\gamma_{(k-1)0g}$) in level 2 as outcome variables. In the following specification of the model, the item and the latent trait effects are specified as random across the courses:

$$\begin{cases} \gamma_{00g} = \alpha_{000} + \xi_{00g} \\ \gamma_{10g} = \alpha_{100} + \xi_{10g} \\ \gamma_{20g} = \alpha_{200} + \xi_{20g} \\ \vdots \\ \gamma_{(k-1)0g} = \alpha_{(k-1)00} + \xi_{(k-1)0g} \end{cases} \tag{4.20}$$

Variation in ξ_{00g} (the satisfaction level estimate for courses) is assumed to be distributed $N(0, \sigma^2_\xi)$ representing the variation among courses in satisfaction.

For item i, combining 4.19 and 4.20 in 4.18 we obtain the log odds for each category:

$$log\left[\frac{p(Y_{ijg} \leq m|x_{ijg}, \beta_{0jg})}{1 - p(Y_{ijg} \leq m|x_{ijg}\beta_{0jg})}\right] = \delta_m - (\alpha_{000} + \xi_{00g} + \mu_{0jg} + \alpha_{i00} + \xi_{i0g} + \mu_{ijg}) \tag{4.21}$$

Equivalence between Multilevel ordered logit and ordinal IRT models

The ordinal IRT model is:

$$P(y_{ij} = m) = \frac{exp(\theta_j - \pi_{im})}{1 - exp(\theta_j - \pi_{im})} \tag{4.22}$$

where m is the category score, θ_j is the latent trait level for person j, and b_{im} is the category difficulty for category m for item i. The items of our questionnaire have four possible response categories (1, 2, 3, 4), then the item will have three category boundary values, b_{i1}, b_{i2} and b_{i3}. The first category corresponds to the probability of getting a score of 2 or 3 or 4 versus a score of 1, the second category value corresponds to the probability of getting a score of 3 or 4 versus 1 or 2, the third one corresponds to the probability of getting 4 versus 1 or 2 or 3.
To demonstrate the equivalence between the parametrization of the ordinal IRT model and multilevel ordered logit model, 4.22 can be manipulated to obtain the following representation of the first category:

$$log\left[\frac{p(Y_{ijg} \leq 1)}{1 - p(Y_{ijg} \leq 1|}\right] = \pi_{i1} - \theta_j \tag{4.23}$$

Equation 4.22 and 4.23 are equivalent:

$$\pi_{i1} - \theta_j = \delta_1 - (\alpha_{000} + \xi_{00g} + \mu_{0jg} + \alpha_{i00} + \xi_{i0g} + \mu_{ijg}) \tag{4.24}$$

It follows that $\mu_{0jg} + \xi_{00g} + \xi_{i0g} + \mu_{ijg}$ is equivalent to the satisfaction parameter θ_j and $\delta_m - (\alpha_{000} + \alpha_{100})$ corresponds to the ability parameter π_i.
Similarly, the value associated with the second category can be written as

$$log\left[\frac{p(Y_{ijg} \leq 2)}{1 - p(Y_{ijg} \leq 2|}\right] = \pi_{i2} - \theta_j \tag{4.25}$$

The equivalence between the models resulting in

$$\pi_{i2} - \theta_j = \delta_2 - (\alpha_{000} + \xi_{00g} + \mu_{0jg} + \alpha_{i00} + \xi_{i0g} + \mu_{ijg}) \tag{4.26}$$

We can obtain same results for the third category.

4.6.3 Adding a student level variable

One of the aim of our analysis is to assess whether some student characteristics affect students satisfaction and so whether student characteristics affect the quality of teaching. In the model described above we can add a student level variable to explain possible differences in satisfaction. We consider for example gender variable. The level 2 model takes the following form:

$$\begin{cases} \beta_{0jg} = \gamma_{00g} + \gamma_{01g}(gender) + \mu_{0jg} \\ \beta_{1jg} = \gamma_{10g} + \mu_{1jg} \\ \beta_{2jg} = \gamma_{20g} + \mu_{2jg} \\ \vdots \\ \beta_{pjg} = \gamma_{(k-1)0g} + \mu_{(k-1)jg} \end{cases} \tag{4.27}$$

and the level 3 model becomes

$$\begin{cases} \gamma_{00g} = \alpha_{000} + \xi_{00g} \\ \gamma_{01g} = \alpha_{010} \\ \gamma_{10g} = \alpha_{100} + \xi_{10g} \\ \gamma_{20g} = \alpha_{200} + \xi_{20g} \\ \vdots \\ \gamma_{(k-1)0g} = \alpha_{(k-1)00} + \xi_{(k-1)0g} \end{cases} \tag{4.28}$$

This model includes the coefficient for gender α_{010}. To obtain the log odds of different categories it is necessary to add this coefficient to the previous formulation in Section 4.6.2.

4.7 Parameter Recovery Study

This simulation study is intended to show parameter recovery for the equivalence between the Multilevel ordered logit model and the ordinal IRT model. In this simulation study, we replicate the data analysis 50 times for the same condition so that we would be able to argue whether multilevel IRT model reproduces ordinal IRT model parameter values. The variables of interest in this simulation study are: sample size ($n = 500$, $n = 1000$), the numbers of items ($k = 10$, $k = 20$). For each replication in each of the four conditions person's satisfaction values are sampled from a standard normal distribution $N(0, 1)$. Item difficulty parameter values are determined so that values are uniformly spaced when items are ordered by quality. Then along with the sampled person-parameter values, the answer probability for the different category of answers is computed for each person by the ordinal Rasch model. Then, the probability value is compared with a random number sampled from a uniform distribution with a range

between 0 and 1. A simulated response is scored 1 (definitely no) if the probability of answer 1 was greater than or equal to the sampled uniform number; the response is scored 2 (more no then yes) if the sampled uniform number is between probability of answer 1 and 2; the response is 3 (more yes than no) for uniform number between probability of 2 and 3; the response is 4 (definitely yes) for uniform between probability of answer 3. The generated data set is analyzed and item and person-parameter are estimated throught a multilevel models. Estimated parameter values are compared across four conditions using mean of correlation coefficient between estimated and true item-parameter values and standard deviations of correlation coefficient. Table 4.1 shows summary statistics from the conditions of the simulation experiment.

The means of correlation coefficients between true and estimated item quality are shown in the third column. The values are consistently very high, greater than 0.98, and they are only different in their third decimal place. Also, their standard deviations, shown in the fourth column, are very small, and they are also only different in their third decimal place. These results show that the reformulated model is able to reproduce item parameter values very well across all the conditions.

Item	sample	mean(r)	sd(r)
10	500	0.985	0.002
	1000	0.989	0.0009
20	500	0.983	0.001
	1000	0.987	0.0007

Table 4.1: Results of parameter recovery study

Chapter 5

Multilevel Rasch model results

This chapter we present the main results for the models introduced in Chapter 4
The first results are related to the PCM then, using satisfaction person parameters of PCM, we apply two-steps Multilevel model and show results. Finally we exhibit the output of the multilevel one-step model for item variables and students' characteristic variables.

5.1 Partial Credit Model results

For select the best model, we can consider some fit statistics. In particular the *Item-trait interaction* test (that approximates a χ^2 distribution) measures the coherence of items. In our case items have a different quality in relation to the lower or greater student satisfaction. In fact $X^2 = 3479.451$.
By the π_i coefficients related to the items we can obtain two important results: calibrate the questionnaires and rank the attribute from the one with the best quality to the least. The observed *misfit* can be decomposed into

contributions of individual items through the analysis of individual parameter estimates, *individual item-fit*. This allows you to identify those items that affect the fit to the model and that, therefore, must be eliminated.
To calibrate the questionnaires we can observe from the output If some quality item parameters which cannot fit correctly.
Table 5.1 shows the item location parameter π_i, the values with the corresponding p.value, the misfit values. Item B5, B6, B7, B11, D2, D3, E1 have to be deleted.

Item	Chisq	df	p-value	Outfit MSQ	Infit MSQ
B3	6239.229	9146	1.000	0.682	0.696
B4	7908.201	9179	1.000	0.862	0.860
B5	12558.625	9037	0.000	1.390	1.269 *
B6	9799.067	9209	0.000	1.064	1.051 *
B7	10299.168	9127	0.000	1.128	1.095 *
B8	7852.911	9231	1.000	0.851	0.851
B10	9408.018	9201	0.064	1.022	1.015
B11	10010.667	9175	0.000	1.091	1.056 *
C1	8006.306	9253	1.000	0.865	0.904
C2	5876.676	9236	1.000	0.636	0.642
D1	8888.761	9227	0.994	0.963	0.966
D2	9565.748	9194	0.003	1.040	1.035 *
D3	10480.979	9122	0.000	1.149	1.125 *
E1	10773.069	9249	0.000	1.165	1.109 *
F2	8793.351	9132	0.994	0.963	0.977
F3	7120.461	9176	1.000	0.776	0.823
F4	6906.089	8855	1.000	0.780	0.814
F5	5895.798	9173	1.000	0.643	0.726
F6	6094.177	9193	1.000	0.663	0.698
F7	5684.363	9181	1.000	0.619	0.645

Table 5.1: Fit statistics of initial model

The best model is reduced to 7 items (Table 5.2. The possible causes of misfit are different: B5, B10, B11 are items that relate to aspects (load of study, teaching coordination) probably require a general knowledge that

the student has not yet, B6, B7 (aspects related to the evaluation of the activities) are not present in all degree courses, D2, D3 (organization) do not concern the teaching, E1 (infrastructure) does not affect the quality of the work by the teacher.

If we sort the items by quality parameter we obtain a ranking of the items from the one with the best quality to the one with the least (Table 5.2), according to the interpretation of the scale given in the previous Chapter. In Table 5.2 we can see the quality items values and threshold parameters. All items in this table refer to the teacher (availability to clarification, observed school hours and the timetable of receiving, educational objectives, clarity). Items F6 (teacher motivation) and C2 (overall student satisfaction) represent educational aspects for whom students perceive quality levels of education lower.

Item	π_i	Threshold 1	Threshold 2	Threshold 3
F5	-0.325	-1.227	-0.895	1.148
F3	-0.285	-1.303	-0.817	1.268
F4	-0.111	-0.963	-0.766	2.062
B3	-0.723	-0.851	-0.281	2.738
F7	-0.961	-0.097	-0.400	2.580
F6	1.048	-0.142	-0.482	2.803
C2	1.210	-0.247	-0.644	3.234

Table 5.2: Fit statistics of final model

In figure 5.1 the Category Probability Curves are plotted for all items. In the horizontal axes we put the person satisfaction values and in the vertical axis the probability related to each response category. We can observe

that for items F3 and F5, the higher category of response are more probable than other items. Moreover for items F4 and F5 there are quasi perfect overlapping between the first and the second category. This points to a possible bad choice of the number of categories.

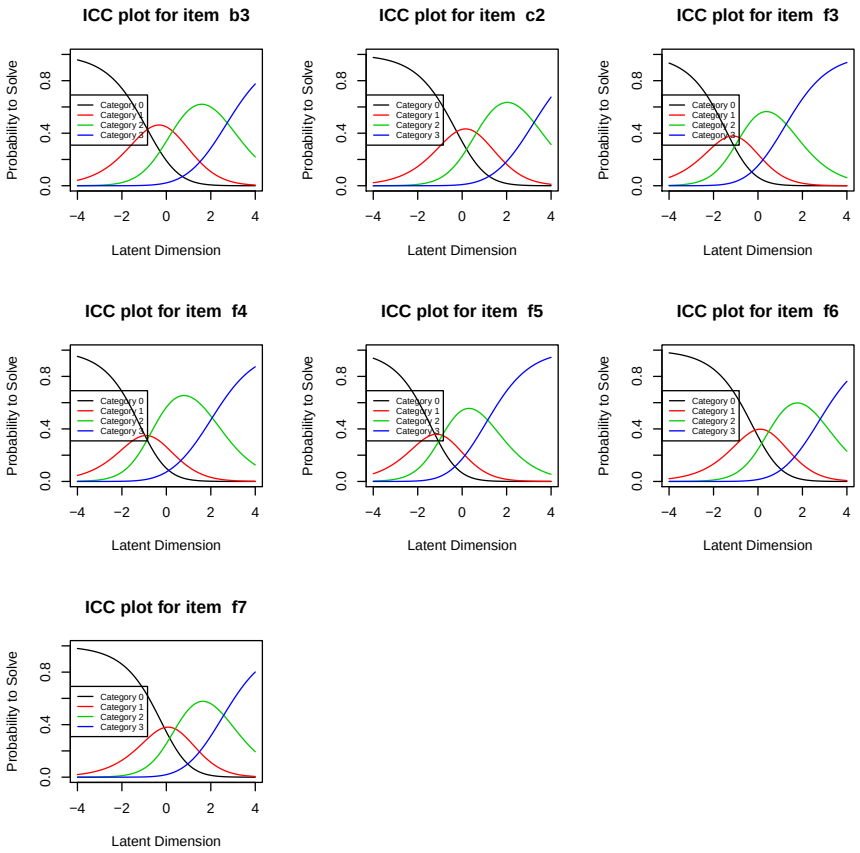

Figure 5.1: Category Probability Curves for all items

If we sort the satisfaction parameters, we obtain a ranking of the sub-

jects from the one least satisfied to the most satisfied. The satisfaction parameters are useful to segment a population with the aim of obtaining different clusters of satisfaction. In our case we can observe (Table 5.3) from LRT (LIkelihood Ratio Test (Andersen, 1973) that item score is different (DIF - Differential Item Functioning) from students who attend high school than students who attend other schools; item score is different from bad and good students ($ISP*$); there is a different level of satisfaction among students of different ages; there is no difference between female and male students.

LRT	p-value	DIF
Gender	0.203	no
School	0.006	yes
Age	0.045	yes
ISP*	0.031	yes

Table 5.3: LRT test

5.1.1 Levels of satisfaction and quality of teaching

The relationship between the location of the item and location of persons along the continuum can be detected by the Figure 5.2 in which the Item/Person Threshold Distribution is shown. The large difference between the standard deviation of the subjects (1.35) and the item (0.67) indicates that the level of satisfaction expressed is not fully captured by the items. The average location of the parameters relating to the subjects (1.78), also indicates a level of overall satisfaction expressed by the above average quality items (0.41). So, the two scales have not a similar range, this does not guarantee that

there is an equilibrium between quality factor and satisfaction factor. probably selected items fail to capture all levels of satisfaction. This could lead to a reflection on the choice of the items that make up the questionnaire.

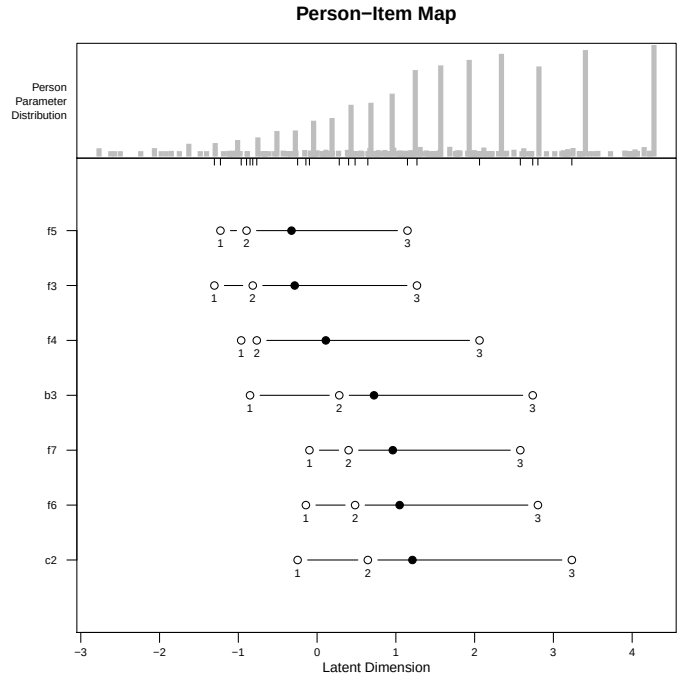

Figure 5.2: Person-Item map

5.2 Multilevel two-step analysis results

In this section we show results from a Multilevel two-step analysis described in Chapter 4.

The first step concerns the estimate of the PCM given in the previous sec-

tion. In the second step we introduce in the multilevel model as response variable the measures of students' satisfaction (satisfaction parameters) obtained by the PCM model in order to detect individual determinants of the level of global satisfaction.

The parameters of the multilevel random intercept model where estimated by *lmer* functionc in **lme4** package.

As previous analysis we consider some characteristics of the student: *age*, *gender*, *school* and $ISP*$. On the first level we can find the students, on the second level we consider the courses. Table 5.4 reports the parameter estimates for both empty model and complete model with students variables.

From this table it follows that the estimate of the grand mean γ_{00} is 1.65. This mean should be interpreted as the expected value of the level of satisfaction for a random student in a randomly drawn course.

The variance between students within the courses about the true course mean is $1.158^2 = 1.341$ (σ^2_ϵ), while the between-group variance (variance between the courses) is $0.711^2 = 0.505$ (σ^2_μ). These variance component estimates give an intraclass correlation coefficient estimate of $\hat{\rho} = 0.505/(0.505 + 1.341) = 0.273$ indicating that about 27% of the variance in satisfaction is between courses.

	Empty model			Model 1		
Fixed	Estim	Std. Err.	p-value	Estim	Std. Err.	p-value
Intercept(γ_{00})	1.650	0.004	0.000	2.951	0.736	0.000
genderM(γ_{10})				-0.110	0.035	0.002
age$\leq$ 22(γ_{20})				-0.022	0.025	0.386
school-				-0.012	0.031	0.712
high(γ_{30})						
ISP*(γ_{40})				-0.884	0.309	0.004
Random	Var.	Std Dev.		Var.	Std Dev.	
σ^2_μ	0.505	0.711		0.515	0.718	
σ^2_ϵ	1.341	1.158		1.327	1.152	

Table 5.4: Parameter estimates: multilvel two-step analysis

To explain variance at the individual level, four level 1 explanatory variables are introduced. The complete model is the following:

$$\begin{cases} y_{ij} = \beta_{0j} + \beta_{1j}(gender) + \beta_{2j}(age) + \beta_{3j}(school) + \beta_{4j}(ISP*) + \epsilon_{ij} \\ \beta_{0j} = \gamma_{00} + \mu_{0j} \\ \beta_{1j} = \gamma_{10} \\ \beta_{2j} = \gamma_{20} \\ \beta_{3j} = \gamma_{30} \\ \beta_{4j} = \gamma_{40} \end{cases} \tag{5.1}$$

The grand mean is γ_{00} and the regression coefficients for *gender*, *age*, *school* and *ISP** are γ_{10}, γ_{20}, γ_{30} and γ_{40}. The random effects μ_{0j} are the level-2 residuals, controlling for the effects of variables. From output (Table 5.4) we can show that the overall level of satisfaction (2.951) has increased if compared to the empty model. Moreover, it can be noted that

age and school effects are not significant. Males are less satisfied then females; *good* students are less satisfied than *bad* students. Let us make a remark: the effect of *age* is not significant, while the effect of ISP^* is significant. Probably this is due to the correlation between these two variables. In fact, we recall that ISP^* is function of *age*, and then could capture the significance of performance and age.

5.3 Multilevel one-step analysis results

In this section we show the results of the three-level model, with two levels of aggregation: students nested in courses. To take into account the variability between items, between students and between courses, it is decided to include random effects for each of these variables. Initially, we estimated a model without explanatory variables with only random effects (Table 5.5). Then we considered a model with the items (Table 5.6).

Random	Var	Std. Dev.
item ($\sigma_{\mu_{ijg}}$)	0.802	0.895
student ($\sigma_{\mu_{0jg}}$)	1.023	1.011
course ($\sigma_{\mu_{00g}}$)	0.435	0.659

Table 5.5: Empty model

Coefficients	Estim.	Std.Err	p.value
itemb11(α_{100})	-0.124	0.0309	0.000
itemb3 (α_{200})	0.724	0.0309	0.000
itemb4 (α_{300})	0.626	0.0313	0.000
itemb8 (α_{400})	0.254	0.0307	0.000
itemc2 (α_{500})	0.278	0.0304	0.000
itemd1 (α_{600})	-0.466	0.0307	0.000
itemd2 (α_{700})	-1.259	0.0309	0.000
itemd3 (α_{800})	0.067	0.0315	0.031
itemf3 (α_{900})	1.989	0.0335	0.000
itemf4 (α_{1000})	1.420	0.0336	0.000
itemf5 (α_{1100})	2.089	0.0309	0.000
itemf6 (α_{1200})	0.503	0.0309	0.000
itemf7 (α_{1300})	0.644	0.0311	0.000

Table 5.6: Model with items

We can see that the items are all significant, except item D3. In particular, the items for which students are more satisfied than the reference item B10, are the items B3, B4, B8, C2, D3, F3, F4, F5, F6, F7, whose coefficients have positive values. In particular, F5 (2.086), F3 (1987), B3 (0.723), F7 (0.637) are the items for which there is greater satisfaction. These items seem to lead to high levels of quality of teaching.

The item D2 is less satisfactory. In the third estimated model we introduce the explanatory variables *sex*, *age* and *school* (Table 5.7). At this moment we don't include the variable ISP^*.

Random	Var	Std. Dev.	
item ($\sigma_{\mu_{ijg}}$)	0.798	0.893	
student ($\sigma_{\mu_{0jg}}$)	1.007	1.004	
course ($\sigma_{\mu_{00g}}$)	0.434	0.658	
Coefficients	**Estim.**	**Std.Err**	**p.value**
itemb11 (α_{100})	-0.129	0.0319	0.000
itemb3 (α_{200})	0.723	0.0319	0.000
itemb4 (α_{300})	0.615	0.0323	0.000
itemb8 (α_{400})	0.251	0.0317	0.000
itemc2 (α_{500})	0.277	0.0314	0.000
itemd1 (α_{600})	-0.469	0.0317	0.000
itemd2 (α_{700})	-1.269	0.0319	0.000
itemd3 (α_{800})	0.050	0.0326	0.123
itemf3 (α_{900})	1.987	0.0346	0.000
itemf4 (α_{1000})	1.424	0.0333	0.000
itemf5 (α_{1100})	2.086	0.0345	0.000
itemf6 (α_{1200})	0.498	0.0319	0.000
itemf7 (α_{1300})	0.637	0.0320	0.000
genderM (α_{010})	-0.075	0.0341	0.027
schoolhigh (α_{020})	-0.013	0.0298	0.653
age$\leq$ 22 (α_{030})	0.117	0.0430	0.007

Table 5.7: Model with students characteristics

The variances of the random effects decrease slightly. We can see that the sex and age characteristics of students influence their satisfaction. In particular, younger students are more satisfied and males are less satisfied

than women.

The random effects μ_{ijg}, μ_{0jg} and ξ_{00g} are not parameters, so they cannot be estimated in the conventional sense, but a "best guess"is provided by the conditional modes. Similarly the conditional variances provides an uncertainty measure of the conditional modes. In Figure 5.3 we present normal probability plots of the conditional modes of the random effects for the each factor.

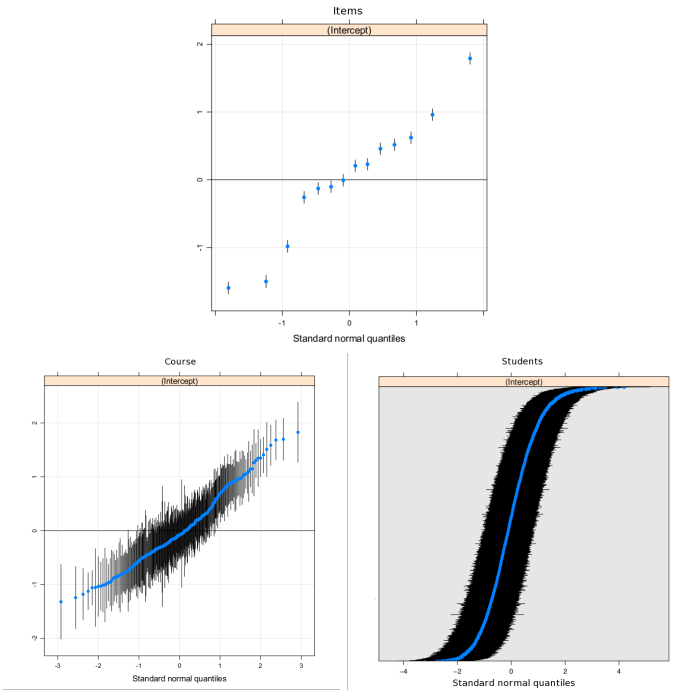

Figure 5.3: Category Probability Curves for all items

To provide a measure of the precision of the conditional distribution of these random effects we add lines extending ±1.96 conditional standard deviations in each direction from the plotted point. We can see that many of

the intervals, above all for students and courses, overlap with the zero line but there are several levels that are clearly greater than zero or clearly less than zero. As indicated by the estimates of the variances of the random effects, the student factor accounts for the greatest level of variability. There are students with low levels of satisfaction, and students with high levels. So it seems that students perceive the quality of teaching in different ways. Table 5.8 show answer probabilities for some items and different categories, considering different profiles of students. It can be see that for all items the probabilities of answer *definitely yes* is highest for women aged less than 22. The probabilities for the first category are similar. For item B3, B8, F3, F6, F7 the highest probability of answer *more no than yes* refers to males aged less than 22. Comparing males and females, regardless their age, women are increasingly satisfied for items B3, B4, B8, F3, F6 and F7. These results highlight, in more detail, the differences in terms of satisfaction among students with different characteristics and so the different perception of quality that students, with different characteristics, have towards facilities provided to them.

Item	gender	age	m	p_m
B3	M	≤ 22	1	0.042
			2	0.147
			3	0.519
			4	0.291
	M	> 22	1	0.047
			2	0.161
			3	0.523
			4	0.268
	F	≤ 22	1	0.039
			2	0.139
			3	0.514
			4	0.346
	F	> 22	1	0.044
			2	0.152
			3	0.521
			4	0.327
B4	M	≤ 22	1	0.040
			2	0.142
			3	0.516
			4	0.341
	M	> 22	1	0.045
			2	0.137
			3	0.540
			4	0.322
	F	≤ 22	1	0.037
			2	0.134
			3	0.511
			4	0.355
	F	> 22	1	0.042
			2	0.147
			3	0.519
			4	0.335
B8	M	≤ 22	1	0.042
			2	0.149
			3	0.518
			4	0.336
	M	> 22	1	0.047
			2	0.159
			3	0.523
			4	0.318
	F	≤ 22	1	0.039
			2	0.137
			3	0.513
			4	0.349
	F	> 22	1	0.043
			2	0.150
			3	0.520
			4	0.329

Item	gender	age	m	p_m
F3	M	≤ 22	1	0.041
			2	0.145
			3	0.518
			4	0.337
	M	> 22	1	0.046
			2	0.158
			3	0.523
			4	0.318
	F	≤ 22	1	0.038
			2	0.137
			3	0.513
			4	0.350
	F	> 22	1	0.043
			2	0.149
			3	0.519
			4	0.330
F6	M	≤ 22	1	0.041
			2	0.144
			3	0.517
			4	0.338
	M	> 22	1	0.046
			2	0.157
			3	0.523
			4	0.319
	F	≤ 22	1	0.038
			2	0.136
			3	0.513
			4	0.351
	F	> 22	1	0.043
			2	0.149
			3	0.519
			4	0.331
F7	M	≤ 22	1	0.041
			2	0.142
			3	0.516
			4	0.341
	M	> 22	1	0.045
			2	0.156
			3	0.522
			4	0.322
	F	≤ 22	1	0.038
			2	0.134
			3	0.646
			4	0.354
	F	> 22	1	0.042
			2	0.147
			3	0.519
			4	0.334

Table 5.8: Answer probability for each category

Table 5.9 presents model with the inclusion of ISP* variable. We can see that coefficients for the items are similar than coefficients in Table 5.7, but the significance of the characteristics of students change. The variable school is still not significant, but now age is also non significant. The ISP* variable is significant. The explanation for this is the same as presented in Section 5.3 *ISP** could capture the *age* effect, it seems that *age* and $ISP*$ are highly correlated. Moreover good students are less satisfied than bad students.

Coefficients	Estim.	Std. Err	p.value
itemb11(α_{100})	-0.129	0.0319	0.000
itemb3 (α_{200})	0.723	0.0319	0.000
itemb4 (α_{300})	0.615	0.0324	0.000
itemb8 (α_{400})	0.251	0.0318	0.000
itemc2 (α_{500})	0.276	0.0315	0.000
itemd1 (α_{600})	-0.469	0.0318	0.000
itemd2 (α_{700})	-1.269	0.0320	0.000
itemd3 (α_{800})	0.050	0.0326	0.124
itemf3 (α_{900})	1.988	0.0346	0.000
itemf4 (α_{1000})	1.424	0.0334	0.000
itemf5 (α_{1100})	2.086	0.0347	0.000
itemf6 (α_{1200})	0.498	0.0319	0.000
itemf7 (α_{1300})	0.637	0.0320	0.000
schoolhigh (α_{010})	-0.002	0.0118	0.859
age$\leq$ 22 (α_{020})	-0.029	0.0586	0.614
genderM (α_{030})	-0.082	0.0336	0.014
ISP* (α_{040})	-0.568	0.1558	0.000

Table 5.9: Model with ISP* variable

These results are similar to results seen in multilevel two-steps analysis. But, in this model we have results for items and students characteristics simultaneously; while in multilevel two-step model we obtain items results from PCM. Moreover, in multilevel two-step models, the standard error of $ISP*$ is twice than standard error in multilevel one step-analysis. This may highlights the drawbacks of multilevel two-step model discussed in Chap-

ter 3.

The next tables (Table 5.10 and 5.11) show the values of quality item parameters differentiated by *gender* and *ISP**. Item F3, F4 and F5 show high quality, in particular females with a bad performance have greater levels of quality than other students profiles. For other items males with a good career have less levels quality than students with other characteristics. Low quality for items B11, D1 and D2 highlights that students are not satisfied of organizational aspects of courses. To sum up, it seems that some items are the drivers of the teaching quality, with some difference between students with different characteristics.

Item	Gender	ISP*	quality item parameter
B11	M	good	-0.780
	M	bad	-0.211
	F	good	-0.697
	F	bad	-0.129
B3	M	good	0.072
	M	bad	0.641
	F	good	0.155
	F	bad	0.723
B4	M	good	-0.036
	M	bad	0.532
	F	good	0.046
	F	bad	0.615
B8	M	good	-0.400
	M	bad	0.168
	F	good	-0.318
	F	bad	0.251
C2	M	good	-0.374
	M	bad	0.194
	F	good	-0.292
	F	bad	0.277
D1	M	good	-1.12
	M	bad	-0.552
	F	good	-1.038
	F	bad	-0.469
D2	M	good	-1.920
	M	bad	-1.352
	F	good	-1.838
	F	bad	-1.269

Table 5.10: Quality item values

Item	Gender	ISP*	quality item parameter
F3	M	good	1.337
	M	bad	1.905
	F	good	1.419
	F	bad	1.988
F4	M	good	0.773
	M	bad	1.342
	F	good	0.856
	F	bad	1.424
F5	M	good	1.435
	M	bad	2.004
	F	good	1.518
	F	bad	2.086
F6	M	good	-0.153
	M	bad	0.415
	F	good	-0.071
	F	bad	0.498
F7	M	good	-0.014
	M	bad	0.554
	F	good	0.068
	F	bad	0.637

Table 5.11: Quality item values (continued)

With reference to the results of the PCM the selected items are the same. But if we consider the relative importance of items (Chapter 3), the items drivers of the quality are different from those obtained in the one-step model.

Concluding remarks

The main issue of this research is the difficulty in identifying the most suitable process of measurement of a latent multidimensional construct such as 'quality of teaching activities'. That difficulty is present in all approaches used in this thesis. In the regression context, the approach of the relative importance metrics considers the overall satisfaction item as a proxy of our latent construct. This approach represents an effort to provide policy makers the drivers of satisfaction/quality identified by the students with *good* and *bad* performance. The results are different between the *good* and *bad* students, despite some overlapping in some aspects. For istance, in both groups of students, items B3 (formative objectives explained by the teacher) and F7 (clarity of the teacher) are important in explaining satisfaction, but the intensity is different in the two groups. Moreover, item F6 (the teacher motivates the interest in the subject) explains satisfaction just for *good* students. In other words, it looks like that performance is a "discriminating" indicator when analyzing B3, F6, and F7 items. These results provide a useful starting point to construct a 'quality teaching indicator' based on performance without controlling for students' characteristics.

On the other hand, an alternative way to the previous data may be given by an analysis on individual data in which it is possible to take into ac-

count the characteristics of the students. By contrast, if you want to take into account the characteristics of the students not only synthesized by the ISP^* indicator, an analysis of individual data is the natural study. Moreover, the sharing of the same course, the same teacher and the same class lead to a model that takes into account the hierarchical structure of data and the diversity that characterize the students. We are talking about multilevel model model as a framework for IRT models for ordinal data. The items, regarding the teacher, are the most important ones as measuring the perceived'quality of teaching activities'. In particular, the results for the PCM show low satisfaction for the items C2 and F6, differing satisfaction for *school*, *age*, and *performance*, but not for *gender*. In general, the results highlight level of satisfactions which are not completely captured by the items.

Then, we introduced in the multilevel model the measures of satisfaction, obtained by the PCM (two-step analysis), in order to assess the importance of individual characteristics in explaining the overall satisfaction. The main results are: the effect *school* and *age* are not significant, males are less satisfied than females, the *good* students are less satisfied than the *bad* students. Thus, the perception of 'quality of teaching' seems to be conditioned by the gender and by the performance.

Introducing a three-level model with two levels of aggregation (one-step analysis), in which students are nested into courses, we can observe - forgetting variables related to students characteristics - the items with a greater satisfaction are F5, F3, B3, F7. Again those items are related to the characteristics of the teacher. On the other hand, when we introduce the explanatory variables *gender*, *school*, and ISP^*, the variable ISP^* influences the satisfaction and we can see *good* students are less satisfied than *bad*

students. Moreover, it can be observed the probability of answering 'definitely yes' for the all items is higher for females with less than 22 years. These results are in accordance with PCM results. On the contrary, males less than 22 years have a higher probability to answer 'more no than yes' than females. This reveals a lower perception of 'quality of teaching'.

In summary: the items that seem to be the drivers of high quality teaching are F3, F4 and F5 for both females and males. However this is more marked for females with bad performance. It should be noted that although the value is not very high, the sign for the items and F6, F7 and C2 remains positive for both males and females with a bad performance. For some items the males with a good performance seem less satisfied. The low quality of items B11, D1 and D2 suggest the lack of student satisfaction towards the organizational aspects.

Finally, we want to stress that the chosen approach is strongly affected by the aim: in fact, if the aim is to construct an indicator of the 'quality of teaching' without controlling students characteristics, then the metric PMVD seems appropriate. But, if the objective is to find a measure of the'quality of teaching'in terms of satisfaction and 'quality' contained in each item and the characteristics of the evaluators, the multilevel one-step model seems more informative. Nevertheless, both results do not seem contradictory as the most important items in any approach are those related to the teacher.

Appendix A

The questionnaire

Università degli Studi di Palermo

QUESTIONARIO DI VALUTAZIONE DELLA DIDATTICA

PRIMA DELLA COMPILAZIONE LEGGERE ATTENTAMENTE LE ISTRUZIONI RIPORTATE SUL RETRO

INFORMAZIONI GENERALI

Data di compilazione ...../...../..... Anno accademico: ...../.....

☐ Primo periodo didattico ☐ Secondo periodo didattico

☐ Laurea I° livello N.O. ☐ Laurea specialistica ☐ Laurea V.O. ☐ Laurea a ciclo unico

Denominazione Corso di Studi: .....

Denominazione Insegnamento[1]: .....

A) LO STUDENTE

A1	Età:			≤18 ☐	19 ☐	20 ☐	21 ☐	22 ☐	23 ☐
				24 ☐	25 ☐	26 ☐	27 ☐	>28 ☐	
A2	Sesso	M = maschio	F = femmina	M ☐	F ☐				
A3	Scuola secondaria di provenienza	A = Liceo classico B = Liceo socio-pedagogico C = Ist. tecnico per geometri	D = Liceo scientifico E = Ist. tecnico commerc. F = Ist. professionale	G = Altri licei H = Ist. tecnico industriale I = Altro	A ☐	B ☐	C ☐	D ☐	E ☐	F ☐
					G ☐	H ☐	I ☐			
A4	Residenza	A = In sede	B = Fuori sede pendolare	C = Fuori sede stanziale	A ☐	B ☐	C ☐			
A5	Anno di corso al quale lo studente è iscritto		In corso		1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐
			Ripetente		1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐
			Fuori corso		1 ☐	2 ☐	3 ☐	4 ☐	≥5 ☐	
A6	Numero totale di crediti acquisiti alla data della rilevazione				0-30 ☐	31-60 ☐	61-90 ☐	91-120 ☐	121-150 ☐	151-180 ☐
					181-210 ☐	211-240 ☐	241-270 ☐	271-300 ☐		
A7	Insegnamenti frequentati in questo periodo didattico				1 ☐	2 ☐	3 ☐	4 ☐	≥5 ☐	
A8	Attività lavorativa in questo anno accademico	A = nessuna	B = saltuaria o part-time	C = a tempo pieno	A ☐	B ☐	C ☐			

Legenda: 1 = decisamente no 2 = più no che sì 3 = più sì che no 4 = decisamente sì NA = non applicabile

B) L'INSEGNAMENTO

B1	Quante ore di lezione hai frequentato *(in percentuale)?*	<25% ☐	25-50% ☐	50-75% ☐	>75% ☐	
B2	Quante ore di esercitazioni hai frequentato *(in percentuale)? (se l'insegnamento[1] non prevede esercitazioni, rispondete non applicabile)*	<25% ☐	25-50% ☐	50-75% ☐	>75% ☐	NA ☐
B3	Gli obiettivi formativi dell'insegnamento[1] sono stati illustrati in aula in modo chiaro?	1 ☐	2 ☐	3 ☐	4 ☐	
B4*	Le modalità dell'esame sono state illustrate in aula in modo chiaro?	1 ☐	2 ☐	3 ☐	4 ☐	
B5	L'insegnamento[1] ha contenuti che si sovrappongono a quelli degli altri insegnamenti[1]?	1 ☐	2 ☐	3 ☐	4 ☐	
B6*	Le attività didattiche integrative *(esercitazioni, laboratori, seminari, ecc...)* sono utili ai fini dell'apprendimento? *(se non previste attività didattiche integrative, rispondete non applicabile)*	1 ☐	2 ☐	3 ☐	4 ☐	NA ☐
B7	Le attività didattiche integrative *(esercitazioni, laboratori, seminari, ecc...)* previste all'interno dell'insegnamento[1] sono adeguatamente coordinate tra loro? *(se non previste attività didattiche integrative, rispondete non applicabile)*	1 ☐	2 ☐	3 ☐	4 ☐	NA ☐
B8*	Il materiale didattico *(indicato o fornito)* è adeguato per lo studio della materia?	1 ☐	2 ☐	3 ☐	4 ☐	
B9*	Le conoscenze preliminari possedute sono sufficienti per la comprensione degli argomenti trattati?	1 ☐	2 ☐	3 ☐	4 ☐	
B10*	Il carico di studio richiesto da questo insegnamento[1] è proporzionato ai crediti indicati nel piano di studi?	1 ☐	2 ☐	3 ☐	4 ☐	
B11	L'insegnamento[1] ha contenuti coordinati con altri insegnamenti[1]?	1 ☐	2 ☐	3 ☐	4 ☐	

C) INTERESSE E SODDISFAZIONE

C1*	Sei interessato ai contenuti di questo insegnamento[1]? *(indipendentemente da come è stato svolto)*	1 ☐	2 ☐	3 ☐	4 ☐
C2*	Sei soddisfatto di come è stato svolto questo insegnamento[1]?	1 ☐	2 ☐	3 ☐	4 ☐

D) ORGANIZZAZIONE

D1*	L'organizzazione complessiva *(sedi, orario, esami, ecc..)* degli insegnamenti[1] ufficialmente previsti in questo periodo didattico è accettabile?	1 ☐	2 ☐	3 ☐	4 ☐
D2*	Il carico di studio complessivo degli insegnamenti[1] ufficialmente previsti in questo periodo didattico è sostenibile?	1 ☐	2 ☐	3 ☐	4 ☐
D3	L'orario di svolgimento dell'attività didattica tiene conto dei tempi di spostamento fra le sedi/aule didattiche?	1 ☐	2 ☐	3 ☐	4 ☐

E) INFRASTRUTTURE

E1*	Le aule in cui si svolgono le lezioni sono adeguate? *(si vede, si sente, si trova posto)*	1 ☐	2 ☐	3 ☐	4 ☐	
E2*	I locali e le attrezzature per le attività didattiche integrative *(esercitazioni, laboratori, seminari, ecc..)* sono adeguati? *(se non previste attività didattiche integrative, rispondete non applicabile)*	1 ☐	2 ☐	3 ☐	4 ☐	NA ☐

[1] Per la Facoltà di Medicina e Chirurgia si intende Corso Integrato

Responsabile dell'insegnamento/modulo COGNOME: ..... NOME: .....

Denominazione modulo *(solo se previsto)*: .....

Scrivere in stampatello e per esteso all'interno degli spazi predisposti

Legenda: 1 = decisamente no 2 = più no che si 3 = più si che no 4 = decisamente si NA = non applicabile

F) RESPONSABILE DELL'INSEGNAMENTO/MODULO

F1	Quale percentuale delle ore di lezione tra quelle previste per il docente è stata svolta dal docente stesso?	<50%	50%-80%	>80%		
F2	Nell'impossibilità di svolgere la lezione, il docente avverte con congruo anticipo *(o comunque in tempo utile)?*	1	2	3	4	NA
F3	Il docente rispetta l'orario di svolgimento dell'attività didattica previsto dal calendario o concordato con gli studenti?	1	2	3	4	
F4	Il docente rispetta l'orario previsto per il ricevimento?	1	2	3	4	
F5*	Il docente è disponibile alle richieste di chiarimenti durante le lezioni?	1	2	3	4	
F6*	Il docente stimola/motiva l'interesse verso la disciplina?	1	2	3	4	
F7*	Il docente espone gli argomenti in modo chiaro?	1	2	3	4	

G) DOMANDE RELATIVE AL MODULO (da compilare solo se l'insegnamento è articolato in moduli)

G1	Il modulo ha contenuti che si sovrappongono a quelli degli altri moduli?	1	2	3	4
G2	Le conoscenze preliminari possedute sono sufficienti per la comprensione degli argomenti trattati?	1	2	3	4
G3	Sei interessato ai contenuti di questo modulo? *(indipendentemente da come è stato svolto)*	1	2	3	4
G4	Sei soddisfatto di come è stato svolto questo modulo?	1	2	3	4

* Le domande contrassegnate con un asterisco compongono un questionario minimo che il Comitato Nazionale per la Valutazione del Sistema Universitario e il Consiglio Nazionale degli Studenti Universitari suggeriscono di adottare, al fine di garantire un'omogenea rilevazione su scala nazionale e assicurare la compatibilità dei dati.

(1) Per la Facoltà di Medicina e Chirurgia si intende Corso Integrato

MARCATE LE CASELLE COSÌ: ■ **E NON COSÌ:** ☒ ☑ ▬ ▬ ▭

ISTRUZIONI PER LA COMPILAZIONE

A) Scrivere esclusivamente con una penna nera o blu
B) Scrivere in stampatello e per esteso all'interno degli spazi predisposti
C) Annerire esclusivamente la casella corrispondente alla risposta esatta. Non sono ammesse correzioni di alcun tipo. Saranno considerate errate le risposte per le quali lo studente abbia annerito più caselle o apportato correzioni
D) Non piegare, sgualcire o macchiare il questionario

SI RICORDA CHE I QUESTIONARI COMPILATI SONO RIGOROSAMENTE ANONIMI E LE INFORMAZIONI CONTENUTE SARANNO ELABORATE E DIFFUSE SOLO IN FORMA AGGREGATA

c. 120.000 - Tipografia PETRONE - 091.302612

Appendix B

Items correlation matrix

	B3	B4	B8	B10	B11	C2	D1	D2	D3	E1	F2	F3	F4	F5	F6	F7
B3	1.000	0.697	0.588	0.316	0.322	0.834	0.301	0.301	0.352	0.354	0.324	0.444	0.482	0.677	0.818	0.790
B4		1.000	0.444	0.345	0.387	0.616	0.392	0.354	0.321	0.362	0.363	0.467	0.410	0.547	0.619	0.603
B8			1.000	0.331	0.401	0.652	0.321	0.377	0.335	0.367	0.361	0.384	0.432	0.431	0.594	0.599
B10				1.000	0.321	0.419	0.373	0.440	0.237	0.345	0.300	0.345	0.316	0.387	0.453	0.401
B11					1.000	0.333	0.345	0.367	0.378	0.398	0.345	0.324	0.453	0.467	0.456	0.456
C2						1.000	0.334	0.249	0.312	0.336	0.389	0.461	0.550	0.722	0.877	0.890
D1							1.000	0.694	0.545	0.384	0.345	0.342	0.340	0.234	0.312	0.342
D2								1.000	0.339	0.365	0.299	0.332	0.298	0.311	0.316	0.318
D3									1.000	0.366	0.321	0.345	0.156	0.299	0.303	0.312
E1										1.000	0.321	0.320	0.343	0.306	0.305	0.228
F2											1.000	0.575	0.543	0.522	0.499	0.501
F3												1.000	0.678	0.534	0.677	0.555
F4													1.000	0.589	0.861	0.506
F5														1.000	0.775	0.876
F6															1.000	0.899
F7																1.000

Table B.1: Items correlation matrix

Bibliography

Achen, C. (1982). *Interpreting and Using Regression*. Sage, Thousand Oaks, CA.

Adam, R., Wilson, M., and Wu, M. (1997). Multilevel item response models: An approach to errors in variable regression. *Journal of Educational and Behavioral Statistic*, **22**, 47–76.

Andersen, E. (1973). A goodness of fit test for the rasch model. *Psychometrika*, **38**(1), 123–140.

Andrich, D. (1978). A rating formulation for ordered responses categories. *Psychometrika*, **43**, 561–573.

Battisti, F. D., Nicolini, G., and Salini., S. (2003). The rasch model to measure service quality. Working Paper Dipartimento di Economia e Politica Aziendale-Sezione Statistica e Matematica.

Bernardi, L., Capursi, V., and Librizzi, L. (2004). Measurement awareness: the use of indicators between expectations and opportunities. *SIS: Sezione Specializzata, Atti della XLIII Riunione Scientifica, Bari*.

Bock, R. (1972). Estimating item parameters and latent abilities when responses are scored in two or more categories. *Psychometrika*.

Campostrini, S., Bernardi, L., and Slanzi, D. (2006). Le determinanti della valutazione della didattica attraverso il parere degli studenti. In *VIII International Meeting on Quantitative Methods for Applied Sciences*.

Capursi, V. and Librizzi, L. (2007). *La qualitÃ della didattica: indicatori semplici o composti?* Franco Angeli, Milan.

Centra, J. (1993). *Reflective faculty evaluation*. San Francisco,Jossey-Bass, CA.

Dobson, A. (1983). *Introduction to Statistical Modelling*. Chapmann and Hall, London.

Efron, B. and Tibshrani, R. (1993). Chapmann and Hall, New York.

Feldman, B. (2006). *Using PMVD to understand hedge fund performance drivers*. Risk Books, London.

Feldman, B. (2007). A theory of attribution. In *MPRA Paper 3349*. University Library of Munich, Germany.

Firth, D. (1998). Relative importance of explanatory variables. In *Statistical 28 Issues in the Social Sciences, Stockholm, Oxford: Nuffied College*.

Fischer, G. and Molenaar, I. (1995). *Rasch Models. Foundations, Recent Developments, and Applications*. Spriger-Verlag, New York.

Fox, J. (2001). Multilevel irt using dichotomous and polytomous data. *British Journal of Mathematical and Statistical Psychology*, **58**, 145–172.

Fox, J. and Glas, C. (1991). Bayesia estimation of a multilevel irt model using gibbs sampling. *Psychometrika*, **56**, 589–600.

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology*, **33**, 234–260.

Goldstein, H. (2002). *Multilevel Statistical Models*. Kendall's.

Grömping, U. (2007). Estimators of relative importance in linear regression based on variance decomposition. *The American Statistician*, **61**, 139–147.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, **38**, 79–93.

Kish, L. (1987). *Statistical Design for research*. New Jersey, John Wiley & Sons.

Kruskal, W. (1987a). Relative importance by averaging over orderings. *The America Statistician*, **41**, 6–10.

Kruskal, W. (1987b). Relative importance by averaging over orderings. *The America Statistician*, **41**, 341.

Kulik, J. (2011). Student ratings: validity, utility and controversy. *New Directions for Institutional Research*, **27**, 925.

Leti, G. (1983). *Statistica descrittiva*. Il Mulino, Bologna.

Librizzi, L. (2008). *Una proposta di un indicatore composto della qualitÃ della didattica al 'netto' delle caratteristiche degli studenti*. Ph.D. thesis, Palermo University.

Lindeman, R. (1980). Introduction to bivariate and multivariate analysis. *Scott, Foresman, Glenview, IL.*

Lord, F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: LAwrence Erlbaum Associates, Inc.

Lucadamo (2010). Rasch analysis and multilevel models for the evaluation of the customer satisfaction. *Electronic Journal of Applied Statistical Analysis EJASA, Electron. J. App. Stat*, **3**, 44–51.

Maier, K. (2001). A rasch hierarchical generalized linear model. *Journal of Educational Measurement*, **38**, 79–93.

Mardia, K., Kent, J., and Bibby, J. (1979). Mcgraw-Hill, London, Academic Press.

Marsh, H. (1984). Students' evaluations of university teaching: dimensionality, reliability, validity, potential biases and utility. *Journal of Educational Psychology*, **76**, 707–754.

Marsh, H. (1987). Students' evaluation of univerisity teaching: Research findings methodological issues, and directions for future reseach. *Internation Journal of Educational Research*, **11**, 253–388.

Masters, G. (1982). A rasch model for partial credit score. *Clara*, **47**, 149–174.

Murray, H. (2005). Student evaluation of teaching: Hs it made a difference? In *Annual Meeting of the Society for Teaching and Learning in Higher Education*, pages 1–15. Charlottetown, Prince Edward Island, Canada.

Nardo, M., Saisana, M., Saltelli, A., and Tarantola, S. (2005). Tools for composite indicators building. In *European Commission, EUR 21682 EN, Institute for the Protection and Security of the Citizen, JRC Ispra, Italy*, page 131.

P. de Boeck, M. W. (2004). *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. Springer.

Pagani, L. and Zanarotti, M. C. (2010). Some uses of rasch models parameters in customer satisfaction data analysis. *Quality Technology & Quantitative Management*, **7**, 83–95.

Pastor, D. (2003). The use of multilevel item response theory modeling in applied research: An illustration. *Applied Measurement in Education*, **16**, 223–243.

Patz, R. and Junker, B. (1999). Applications and extension of mcmc in irt: Multiple item types, missing data and rated responses. *Journal of Educational and Behavioral Statistic*, **24**, 342–366.

Rampichini, C. and A. Petrucci, L. G. (2004). Analysis of university course evaluations: from descriptive measures to multilevel models. *Statistical Methods & Applications*, **13**, 357–373.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: MESA PRESS.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Press, Chicago.

Raudenbush, S. and Sampson, R. (1999). Econometrics: Toward a science of assessing ecological settings, with application to the systematic social observation of neighborhoods. *The Annals of Statistics*, **29**, 1–41.

Remmers, H. (1928). The relationship between students' makers and student attituted towards instructors. *School and Society*, **28**, 759–760.

Remmers, H. (1929). Student rating of college teaching. *School and Society*, **30**, 232–234.

Remmers, H. and Brandenburg, G. (1927). Experimental data on the psrdue rating scale for instructors. *Educational Administration and Supervision*, **13**, 519–527.

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. *Psychometrika Monograph*.

Snijders, A. B. and Bosker, R. J. (1999). *An Introduction to Basic Multilevel Modeling*. SAGE Publications, Ltd.

Sulis, I. (2007). *Measuring students' assessment of 'univesity course quality' using mixed-effects models*. Ph.D. thesis, PhD thesis in Applied Statistics.

V. Capursi, L. L. (2008). *La qualitÃ della didattica: indicatori semplici o composti?* Dottor Divago: Discernere valutare e governare la nuova UniverisitÃ . Collana Valuatazione AIV - Teoria, metodologia e ricerca, Franco Angeli. Milano.

Verhelst, N. and Eggen, T. (1989). *IPsychometrics and statistical aspects of measurement research*. Arnhem, The Netherlands: Cito.

Wachtel, H. (1998). Student evaluations of college teaching effectiveness: a brief review. *Assessment and Evalution in Higher Education*, **23**, 191–210.

Wilson, M. (1988). Detecting and interpreting local item dependence using a family of rasch models. *Applied Psychological Measurement*, **21**, 353–364.

Wright, B. and Masters, G. (1982). *Rating Scale Analysis*. MESA Press, Chicago.

Wright, R. (2006). Student evaluation of faculty: Concerns raised in the literature, and posible solutions. *College Student Journal*, **40(2)**, 417–422.

Zwinderman, A. (1991). A generalized rasch model for manifest predictors. *Psychometrika*, **56**, 589–600.

Zwinderman, A. (1997). *Response models with manifest predictors. In W.J. van der Linden & R.K. Hambleton, Handbook of modern item resposne theory*. Springer, New York.